Conditional Adversarial Domain Adaptation

Mingsheng Long[1], Zhangjie Cao[1], Jianmin Wang[1], and Michael I. Jordan[2]

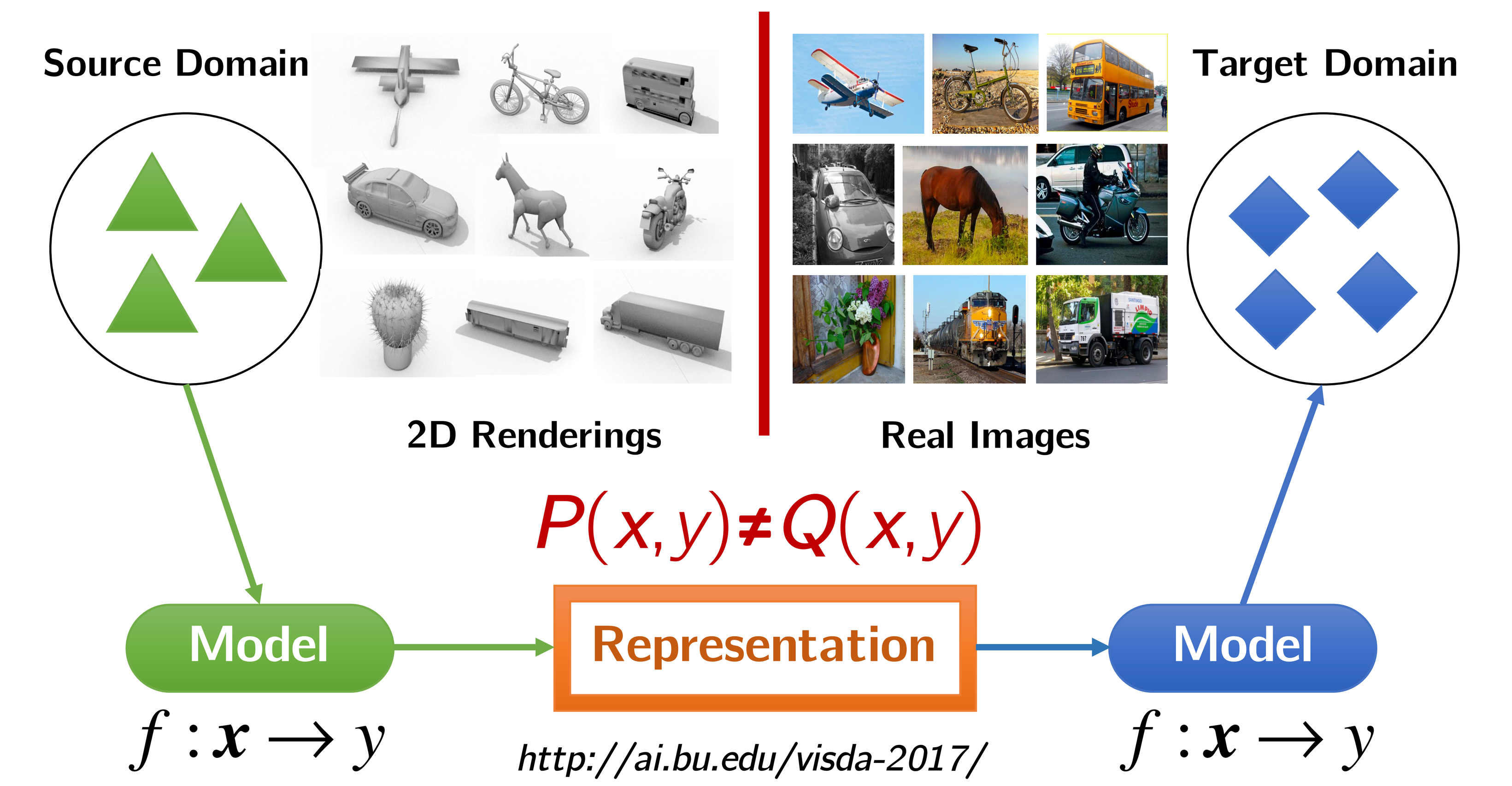
[1]School of Software, Tsinghua University, China [2]Department of EECS, UC Berkeley, USA

Summary

- Principled approaches to domain adaptation: Conditional Domain Adversarial Networks (CDAN)
- Two technical contributions:
  - Multilinear Conditioning: capture the cross-covariance between domain-specific feature representations and classifier predictions to improve the discriminability
  - Entropy Conditioning: control the uncertainty of (target) classifier predictions to guarantee the transferability
- New domain adaptation theory on the generalization error bound
- State-of-art results on many vision & simulation-to-real datasets
- Open Problems
  - Randomized method for multilinear operation with lower approximation error
  - Complexity analysis for the domain adaptation theory involving neural networks
- Code@: `https://github.com/thuml/CDAN`

Deep Domain Adaptation

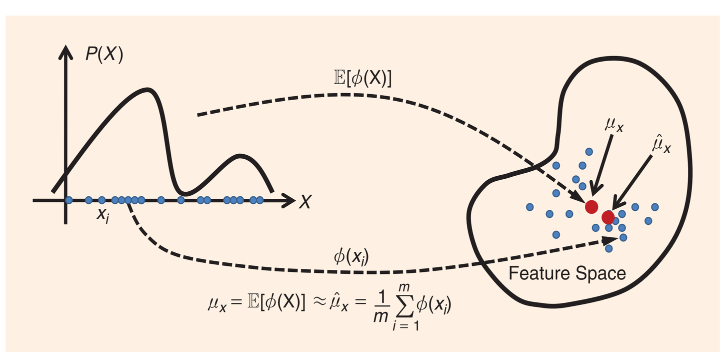
Deep Learning across Domains following Non-IID Distributions $P \neq Q$



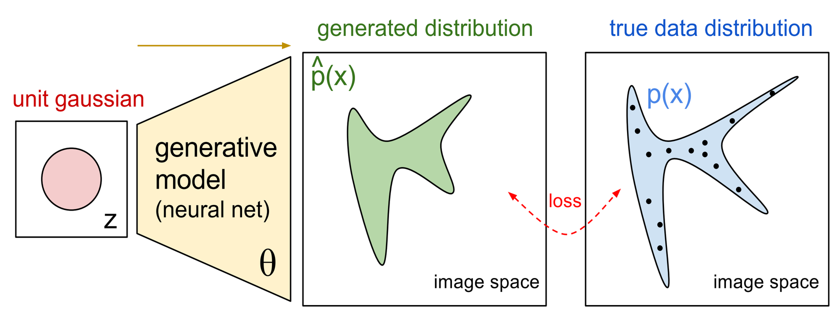
Basic Approaches to Domain Adaptation

Matching distributions across source and target domains s.t. $P \approx Q$

- Reduce marginal distribution mismatch: $P(\mathbf{X}) \neq Q(\mathbf{X})$
- Reduce conditional distribution mismatch: $P(Y|\mathbf{X}) \neq Q(Y|\mathbf{X})$
- Challenge: fail to align different domains of multimodal distributions



Kernel Embedding

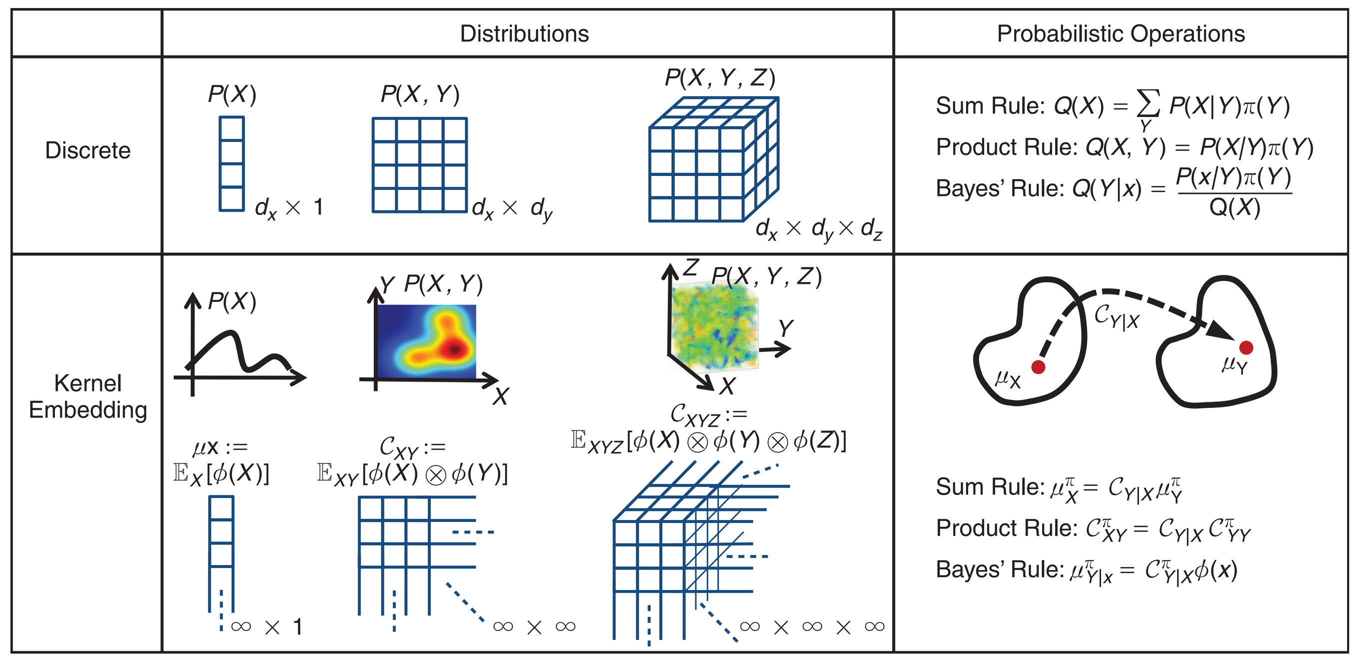


Adversarial Learning

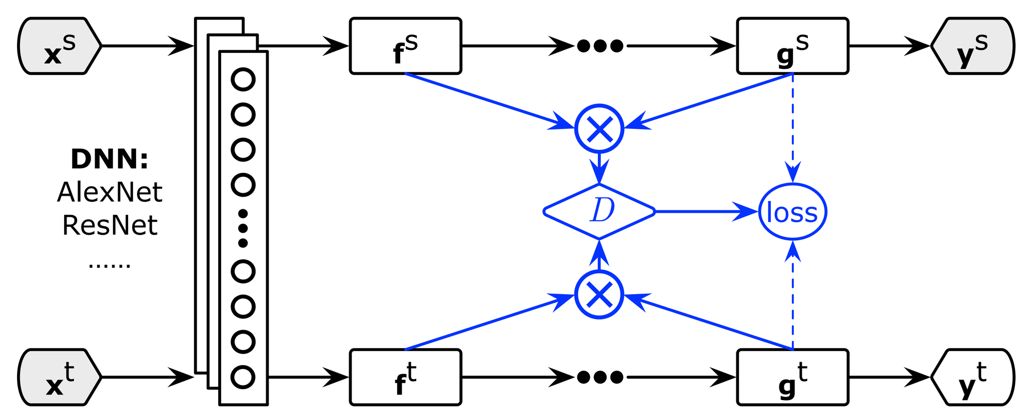
Main Idea of This Work

Distribution Embeddings with Statistics: multilinear ≫ concatenation

- Capture cross-covariance statistics across multiple random vectors



CDAN: Multilinear Conditioning

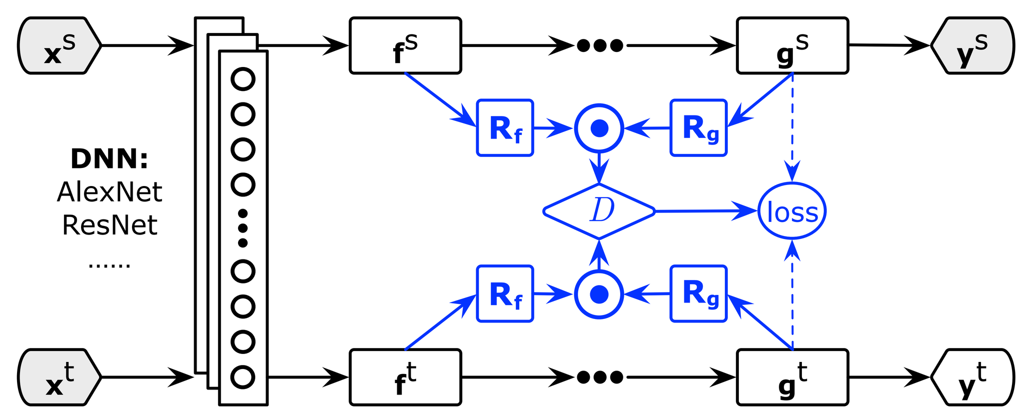


Conditional adaptation of distributions over representation & prediction

$$\min_{G} \; \mathcal{E}(G) - \lambda \mathcal{E}(D, G) \qquad \min_{D} \; \mathcal{E}(D, G), \tag{1}$$

$$\mathcal{E}(D, G) = -\mathbb{E}_{\mathbf{x}_i^s \sim \mathcal{D}_s} \log \left[D\left(\mathbf{f}_i^s \otimes \mathbf{g}_i^s\right)\right] - \mathbb{E}_{\mathbf{x}_j^t \sim \mathcal{D}_t} \log \left[1 - D\left(\mathbf{f}_j^t \otimes \mathbf{g}_j^t\right)\right] \tag{2}$$

CDAN: Randomized Multilinear Conditioning



Conditional adaptation of distributions over representation & prediction

$$T_{\otimes}(\mathbf{f}, \mathbf{g}) = \mathbf{f} \otimes \mathbf{g} \tag{3}$$

$$T_{\odot}(\mathbf{f}, \mathbf{g}) = \frac{1}{\sqrt{d}} (\mathbf{R_f f}) \odot (\mathbf{R_g g}) \tag{4}$$

$$T(\mathbf{h}) = \begin{cases} T_{\otimes}(\mathbf{f}, \mathbf{g}) & \text{if } d_f \times d_g \leqslant 4096 \\ T_{\odot}(\mathbf{f}, \mathbf{g}) & \text{otherwise} \end{cases} \tag{5}$$

CDAN: Entropy Conditioning

$$\begin{aligned} &\min_{G} \mathbb{E}_{(\mathbf{x}_i^s, \mathbf{y}_i^s) \sim \mathcal{D}_s} L\left(G\left(\mathbf{x}_i^s\right), \mathbf{y}_i^s\right) \\ &\quad + \lambda \left(\mathbb{E}_{\mathbf{x}_i^s \sim \mathcal{D}_s} w\left(H\left(\mathbf{g}_i^s\right)\right) \log \left[D\left(T\left(\mathbf{h}_i^s\right)\right)\right] + \mathbb{E}_{\mathbf{x}_j^t \sim \mathcal{D}_t} w\left(H\left(\mathbf{g}_j^t\right)\right) \log \left[1 - D\left(T\left(\mathbf{h}_j^t\right)\right)\right] \right) \\ &\max_{D} \mathbb{E}_{\mathbf{x}_i^s \sim \mathcal{D}_s} w\left(H\left(\mathbf{g}_i^s\right)\right) \log \left[D\left(T\left(\mathbf{h}_i^s\right)\right)\right] + \mathbb{E}_{\mathbf{x}_j^t \sim \mathcal{D}_t} w\left(H\left(\mathbf{g}_j^t\right)\right) \log \left[1 - D\left(T\left(\mathbf{h}_j^t\right)\right)\right] \end{aligned} \tag{6}$$

CDAN: Generalization Error Bound

The probabilistic bound of the target risk $\epsilon_Q(G)$ of hypothesis G is given by the source risk $\epsilon_P(G)$ plus the distribution discrepancy:

$$\epsilon_Q(G) \leqslant \epsilon_P(G) + \left[\epsilon_P(G^*) + \epsilon_Q(G^*)\right] + \left|\epsilon_P(G, G^*) - \epsilon_Q(G, G^*)\right|. \tag{7}$$

The distribution discrepancy $|\epsilon_P(G, G^*) - \epsilon_Q(G, G^*)|$ is bounded by

$$\begin{aligned} \left|\epsilon_P(G, G^*) - \epsilon_Q(G, G^*)\right| &\leqslant \left|\mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim P_G}\left[\mathbf{g} \neq G^*(\mathbf{f})\right] - \mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim Q_G}\left[\mathbf{g} \neq G^*(\mathbf{f})\right]\right| \\ &\leqslant \sup_{G^* \in \mathcal{H}} \left|\mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim P_G}\left[\left|\mathbf{g} - G^*(\mathbf{f})\right| \neq 0\right] - \mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim Q_G}\left[\left|\mathbf{g} - G^*(\mathbf{f})\right| \neq 0\right]\right| \\ &\leqslant \sup_{\delta \in \Delta} \left|\mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim P_G}\left[\delta(\mathbf{f}, \mathbf{g}) \neq 0\right] - \mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim Q_G}\left[\delta(\mathbf{f}, \mathbf{g}) \neq 0\right]\right| \\ &\leqslant \sup_{D \in \mathcal{H}_D} \left|\mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim P_G}\left[D(\mathbf{f}, \mathbf{g}) \neq 0\right] - \mathbb{E}_{(\mathbf{f},\mathbf{g}) \sim Q_G}\left[D(\mathbf{f}, \mathbf{g}) \neq 0\right]\right|, \end{aligned} \tag{8}$$

i.e., the distribution discrepancy is bounded by domain discriminator.

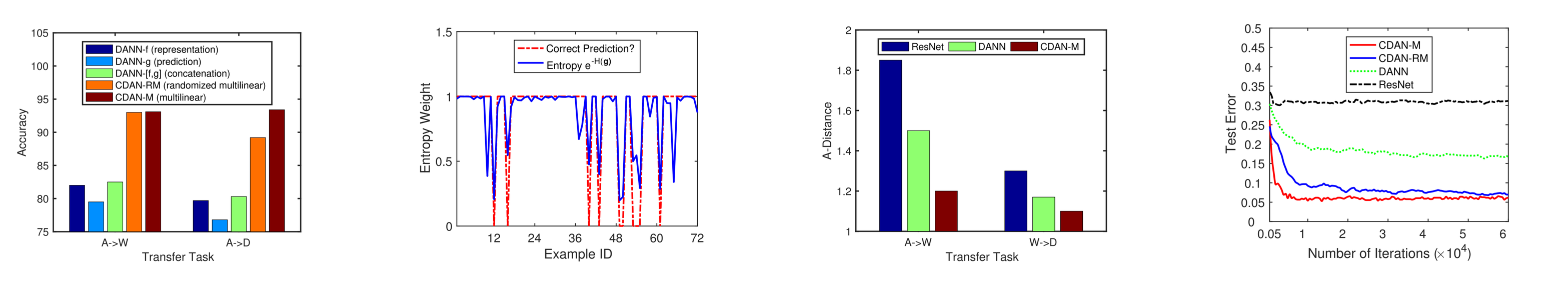
Experimental Results

Table: Accuracy (%) on *Office-31* for Unsupervised Domain Adaptation

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
AlexNet	61.6±0.5	95.4±0.3	99.0±0.2	63.8±0.5	51.1±0.6	49.8±0.4	70.1
DANN	73.0±0.5	96.4±0.3	99.2±0.3	72.3±0.3	53.4±0.4	51.2±0.5	74.3
JAN	74.9±0.3	96.6±0.2	99.5±0.2	71.8±0.2	**58.3**±0.3	55.0±0.4	76.0
CDAN	77.9±0.3	96.9±0.2	**100.0**±.0	75.1±0.2	54.5±0.3	**57.5**±0.4	77.0
CDAN+E	**78.3**±0.2	**97.2**±0.1	**100.0**±.0	**76.3**±0.1	57.3±0.2	57.3±0.3	**77.7**
ResNet-50	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DANN	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
JAN	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
CDAN	93.1±0.2	98.2±0.2	**100.0**±.0	89.8±0.3	70.1±0.4	68.0±0.4	86.6
CDAN+E	**94.1**±0.1	**98.6**±0.1	**100.0**±.0	**92.9**±0.2	71.0±0.3	69.3±0.3	**87.7**

Table: Accuracy (%) on *Office-Home* for Unsupervised Domain Adaptation

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
AlexNet	26.4	32.6	41.3	22.1	41.7	42.1	20.5	20.3	51.1	31.0	27.9	54.9	34.3
DANN	36.4	45.2	54.7	35.2	51.8	55.1	31.6	39.7	59.3	45.7	46.4	65.9	47.3
JAN	35.5	46.1	57.7	36.4	53.3	54.5	33.4	40.3	60.1	45.9	47.4	67.9	48.2
CDAN	36.2	47.3	58.6	37.3	54.4	**58.3**	33.2	**43.9**	62.1	48.2	48.1	70.7	49.9
CDAN+E	**38.1**	**50.3**	**60.3**	**39.7**	**56.4**	57.8	**35.5**	43.1	**63.2**	**48.4**	**48.5**	**71.1**	**51.0**
ResNet-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
CDAN+E	**50.7**	**70.6**	**76.0**	**57.6**	**70.0**	**70.0**	**57.4**	**50.9**	**77.3**	**70.9**	**56.7**	**81.6**	**65.8**



(a) Multilinear (b) Entropy (c) Discrepancy (d) Convergence

Figure: Analysis of conditioning strategies, distribution discrepancy, and convergence.