Partial Transfer Learning with Selective Adversarial Networks

Zhangjie Cao[†], Mingsheng Long[†], Jianmin Wang[†], and Michael I. Jordan[♯]

[†]KLiss, MOE; NEL-BDS; TNList; School of Software, Tsinghua University, China

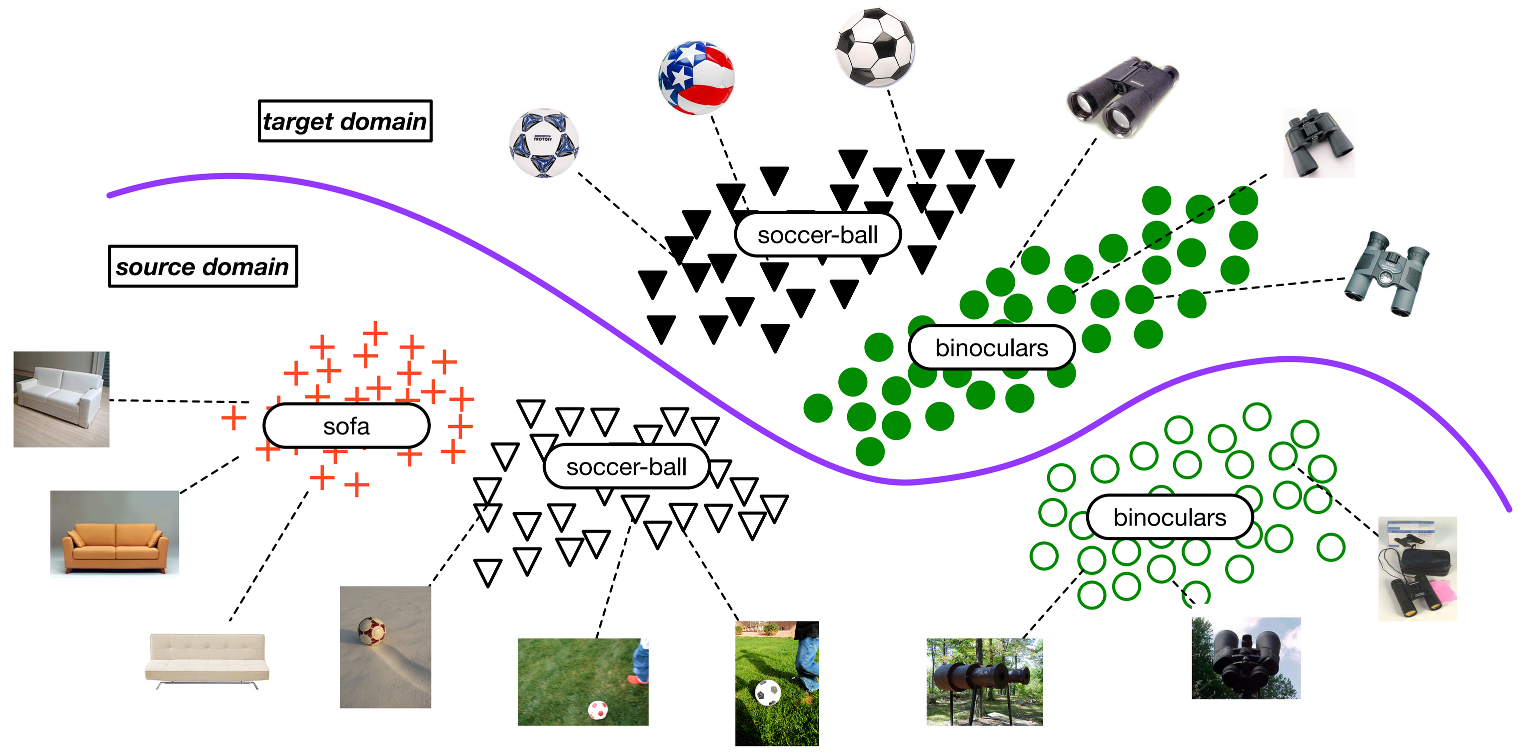
[♯]University of California, Berkeley, Berkeley, USA

Summary

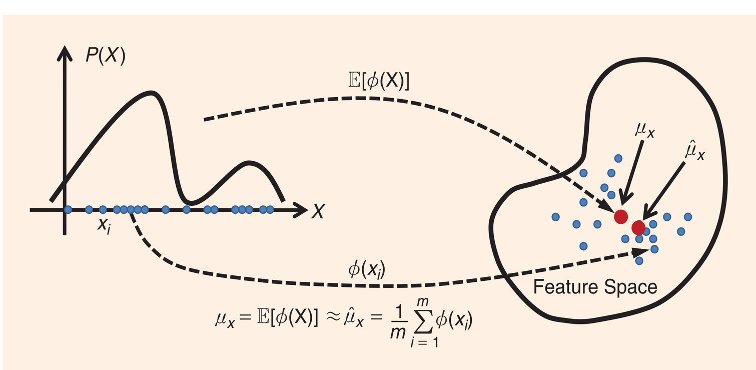
- Partial transfer learning: Deep learning across domains with different label spaces $\mathcal{C}_s \supset \mathcal{C}_t$
- Two main challenges:
  - Positive transfer across domains in shared label space $P_{\mathcal{C}_t} \neq Q_{\mathcal{C}_t}$
  - Negative transfer across domains in outlier label space $P_{\mathcal{C}_s \backslash \mathcal{C}_t} \neq Q_{\mathcal{C}_t}$
- State-of-the-art results on partial transfer learning datasets.
- Main contributions:
  - Propose a multi-adversarial networks architecture to enable class-wise domain distribution matching;
  - Develop a weighting mechanism with instance and class level weight to avoid negative transfer.
- Code available @ `https://github.com/thuml/SAN`

Partial Transfer Learning

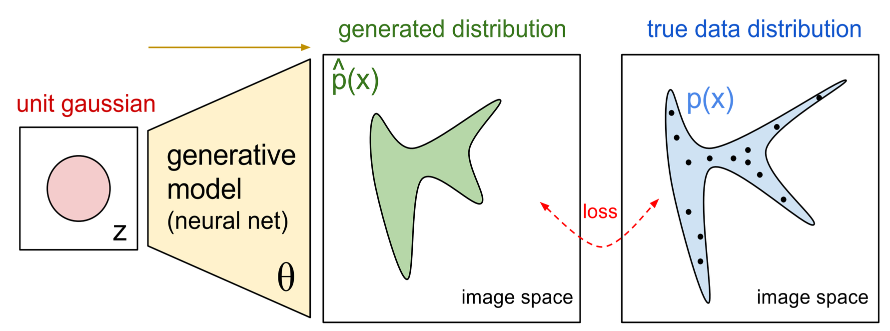
- Deep learning across domains with different label spaces $\mathcal{C}_s \supset \mathcal{C}_t$
- Positive transfer across domains in shared label space $P_{\mathcal{C}_t} \neq Q_{\mathcal{C}_t}$
- Negative transfer across domains in outlier label space $P_{\mathcal{C}_s \backslash \mathcal{C}_t} \neq Q_{\mathcal{C}_t}$



Partial Transfer Learning: How?

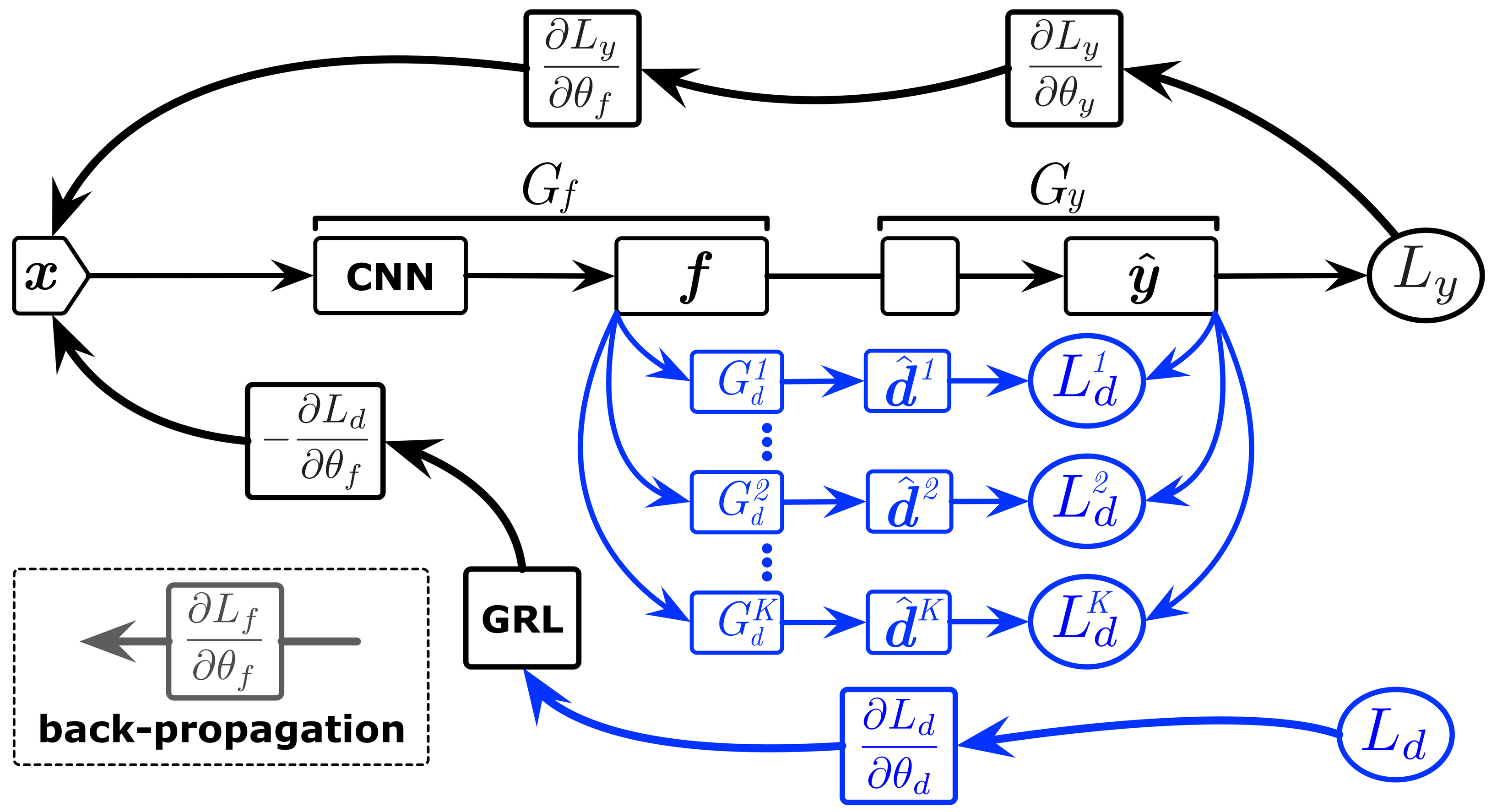


Kernel Embedding



Adversarial Learning

Selective Adversarial Networks



- $\mathbf{f} = G_f(\mathbf{x})$: feature extractor
- $\hat{\mathbf{y}}$: predicted data label
- $\hat{\mathbf{d}}$: predicted domain label
- G_y, L_y: label predictor and loss
- G_d^k, L_d^k: domain discriminator
- GRL: gradient reversal layer

Weighting Mechanism and Loss

- Instance Weighting (IW): probability-weighted loss for $G_d^k, k = 1, \ldots, |\mathcal{C}_s|$.
  Class Weighting (CW): down-weigh $G_d^k, k = 1, \ldots, |\mathcal{C}_s|$ for outlier classes

$$L_d = \frac{1}{n_s + n_t} \sum_{k=1}^{|\mathcal{C}_s|} \left\{ \left(\frac{1}{n_t} \sum_{\mathbf{x}_i \in \mathcal{D}_t} \hat{y}_i^k \right) \times \left(\sum_{\mathbf{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} \hat{y}_i^k L_d^k \left(G_d^k \left(G_f(\mathbf{x}_i) \right), d_i \right) \right) \right\} \quad (1)$$

- Entropy (uncertainty) minimization: $H\left(G_y\left(G_f(\mathbf{x}_i)\right)\right) = -\sum_{k=1}^{|\mathcal{C}_s|} \hat{y}_i^k \log \hat{y}_i^k$

$$E = \frac{1}{n_t} \sum_{\mathbf{x}_i \in \mathcal{D}_t} H\left(G_y\left(G_f(\mathbf{x}_i)\right)\right) \quad (2)$$

- Overall Loss C

$$C\left(\theta_f, \theta_y, \theta_d^k|_{k=1}^{|\mathcal{C}_s|}\right) = \frac{1}{n_s} \sum_{\mathbf{x}_i \in \mathcal{D}_s} L_y\left(G_y\left(G_f(\mathbf{x}_i)\right), y_i\right) + \frac{1}{n_t} \sum_{\mathbf{x}_i \in \mathcal{D}_t} H\left(G_y\left(G_f(\mathbf{x}_i)\right)\right) - \frac{1}{n_s + n_t} \sum_{k=1}^{|\mathcal{C}_s|} \left\{ \left(\frac{1}{n_t} \sum_{\mathbf{x}_i \in \mathcal{D}_t} \hat{y}_i^k \right) \times \left(\sum_{\mathbf{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} \hat{y}_i^k L_d^k \left(G_d^k \left(G_f(\mathbf{x}_i) \right), d_i \right) \right) \right\} \quad (3)$$

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} C\left(\theta_f, \theta_y, \theta_d^k|_{k=1}^{|\mathcal{C}_s|}\right)$$
$$(\hat{\theta}_d^1, \ldots, \hat{\theta}_d^{|\mathcal{C}_s|}) = \arg\max_{\theta_d^1, \ldots, \theta_d^{|\mathcal{C}_s|}} C\left(\theta_f, \theta_y, \theta_d^k|_{k=1}^{|\mathcal{C}_s|}\right) \quad (4)$$

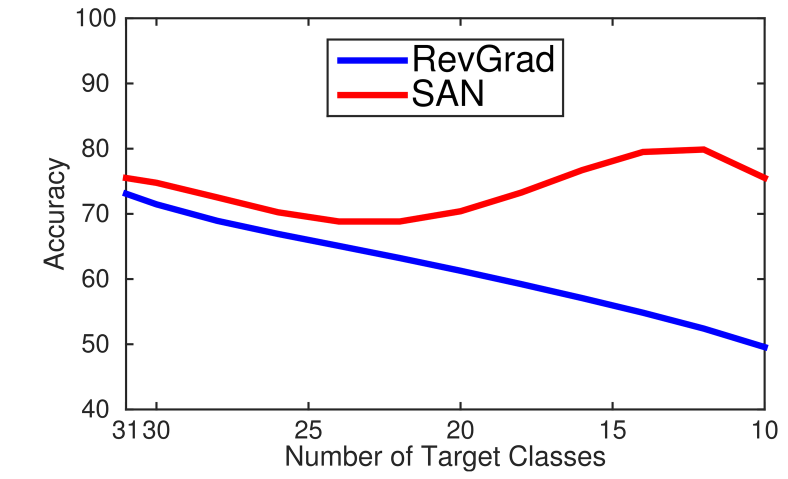
Experimental Results

Table: Accuracy (%) of partial transfer learning tasks on *Office-31*

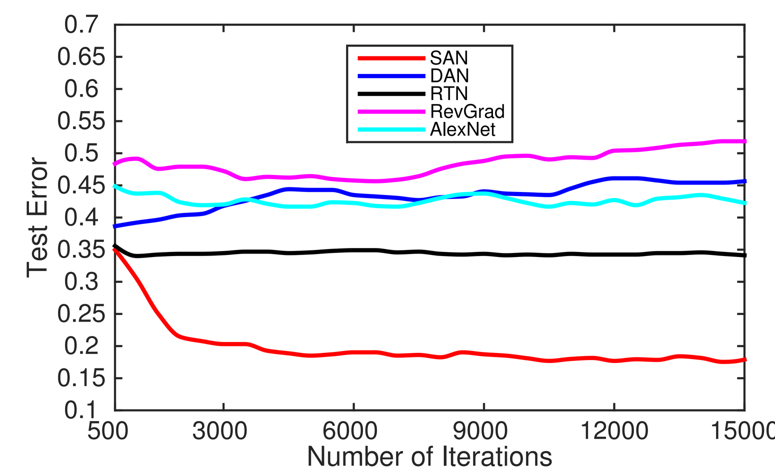
Method	Office-31						
	A 31 → W 10	D 31 → W 10	W 31 → D 10	A 31 → D 10	D 31 → A 10	W 31 → A 10	Avg
AlexNet	58.51	95.05	98.08	71.23	70.6	67.74	76.87
DAN	56.52	71.86	86.78	51.86	50.42	52.29	61.62
RevGrad	49.49	93.55	90.44	49.68	46.72	48.81	63.11
RTN	66.78	86.77	99.36	70.06	73.52	76.41	78.82
ADDA	70.68	96.44	98.65	72.90	74.26	75.56	81.42
SAN-selective	71.51	98.31	100.00	78.34	77.87	76.32	83.73
SAN-entropy	74.61	98.31	100.00	80.29	78.39	82.25	85.64
SAN	**80.02**	**98.64**	**100.00**	**81.28**	**80.58**	**83.09**	**87.27**

Table: Accuracy (%) of partial transfer learning tasks on *Caltech-Office* and *ImageNet-Caltech*

Method	Caltech-Office				ImageNet-Caltech		
	C 256 → W 10	C 256 → A 10	C 256 → D 10	Avg	I 1000 → C 84	C 256 → I 84	Avg
AlexNet	58.44	76.64	65.86	66.98	52.37	47.35	49.86
DAN	42.37	70.75	47.04	53.39	54.21	52.03	53.12
RevGrad	54.57	72.86	57.96	61.80	51.34	47.02	49.18
RTN	71.02	81.32	62.35	71.56	63.69	50.45	57.07
ADDA	73.66	78.35	74.80	75.60	64.20	51.55	57.88
SAN-selective	76.44	81.63	80.25	79.44	66.78	51.25	59.02
SAN-entropy	72.54	78.95	76.43	75.97	55.27	52.31	53.79
SAN	**88.33**	**83.82**	**85.35**	**85.83**	**68.45**	**55.61**	**62.03**

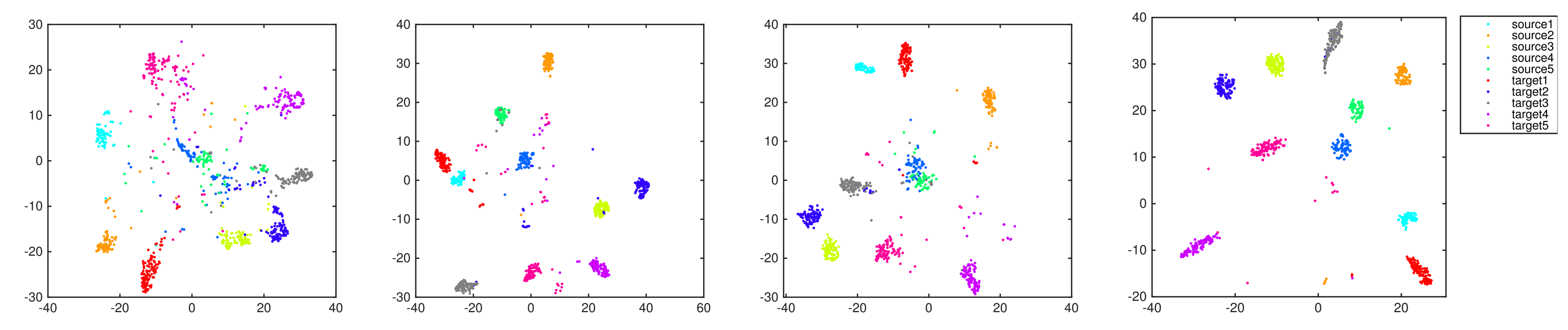


(a) Accuracy w.r.t #Target Classes

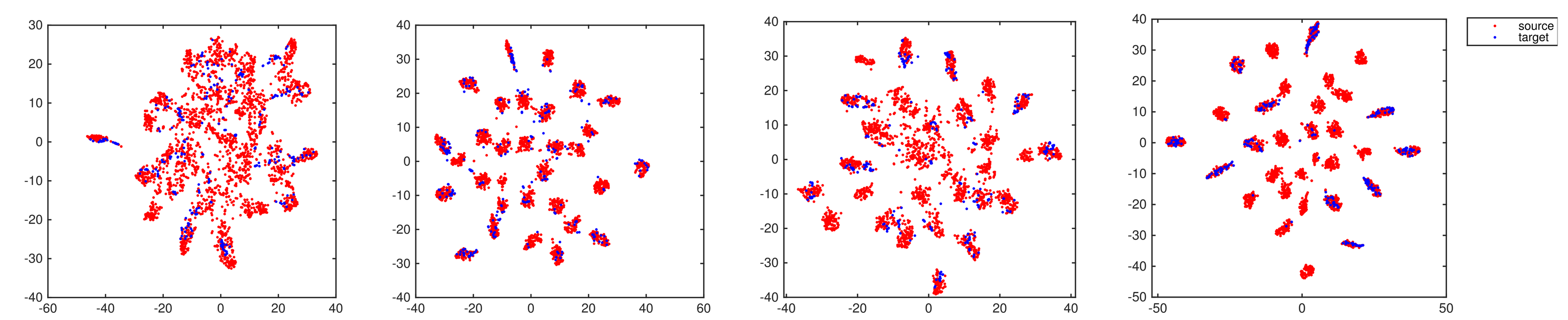


(b) Test Error

Figure: Empirical analysis: (a) Accuracy by varying #target domain classes; (b) Target test error.



(a) DAN (b) RevGrad (c) RTN (d) SAN



(e) DAN (f) RevGrad (g) RTN (h) SAN

Figure: The t-SNE visualization of DAN, RevGrad, RTN, and SAN.