中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

Unsupervised Domain Adaptation with Residual Transfer Networks

Institute of Automation, Chines Academy of Sciences, CASIA
National Laboratory of Pattern Recognition, NLPR
Human Machine Interaction Group
Ye Bai
March 31, 2018

Outline

- Authors
- Motivation
- Methods
- Experiments

Authors

Unsupervised Domain Adaptation with Residual Transfer Networks

Mingsheng Long[†]**, Han Zhu**[†]**, Jianmin Wang**[†]**, and Michael I. Jordan**[♯]
[†]KLiss, MOE; TNList; School of Software, Tsinghua University, China
[♯]University of California, Berkeley, Berkeley, USA
{mingsheng,jimwang}@tsinghua.edu.cn, zhuhan10@gmail.com, jordan@berkeley.edu

龙明盛

Mingsheng Long

Assistant Professor, Ph.D Supervisor
School of Software, Tsinghua University

Machine Learning Group
National Engineering Lab for Big Data Software

Curriculum Vitae
longmingsheng@gmail.com, mingsheng@tsinghua.edu.cn
Room 11-413, East Main Building, Tsinghua University, Beijing, China

1. **Mingsheng Long**, Jianmin Wang, Yue Cao, Jiaguang Sun, Philip S. Yu. **Deep Learning of Transferable Representation for Safe Domain Adaptation**. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 28(8):2027-2040, 2016.
2. **Mingsheng Long**, Jianmin Wang, Jiaguang Sun, Philip S. Yu. **Domain Invariant Transfer Kernel Learning**. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 27(6):1519-1532, 2015.
3. **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Dou Shen, Qiang Yang. **Transfer Learning with Graph Co-Regularization**. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 26(7):1805-1818, 2014.
4. **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, Philip S. Yu. **Adaptation Regularization: A General Framework for Transfer Learning**. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 26(5):1076-1089, 2014.

Outline

- Authors
- Motivation
- Methods
- Experiments

Problem description

- Given
  source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$
  unlabeled target domain $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$
- Source domain and target domain are sampled from different probability distributions.
- How to minimize expected target risk
  $R_t(f_t) = \mathbb{E}_{(\mathbf{x},y)\sim q}[f_t(\mathbf{x}) \neq y]$
  by leveraging the source domain supervised data?

Discrepancy

-

distribution discrepancy

$p(\mathbf{x}, y) \neq q(\mathbf{x}, y)$

feature discrepancy

$p(\mathbf{x}) \neq q(\mathbf{x})$

classifier discrepancy

$f_s(\mathbf{x}) \neq f_t(\mathbf{x})$

no label in target domain

Motivation

- Bridge the source classifier $f_S(\mathbf{x})$ and target classifier $f_T(\mathbf{x})$ by residual layers.

- Model discrepancy as a perturbation function

$$f_S\left(\mathbf{x}\right) = f_T\left(\mathbf{x}\right) + \Delta f\left(\mathbf{x}\right),$$

Outline

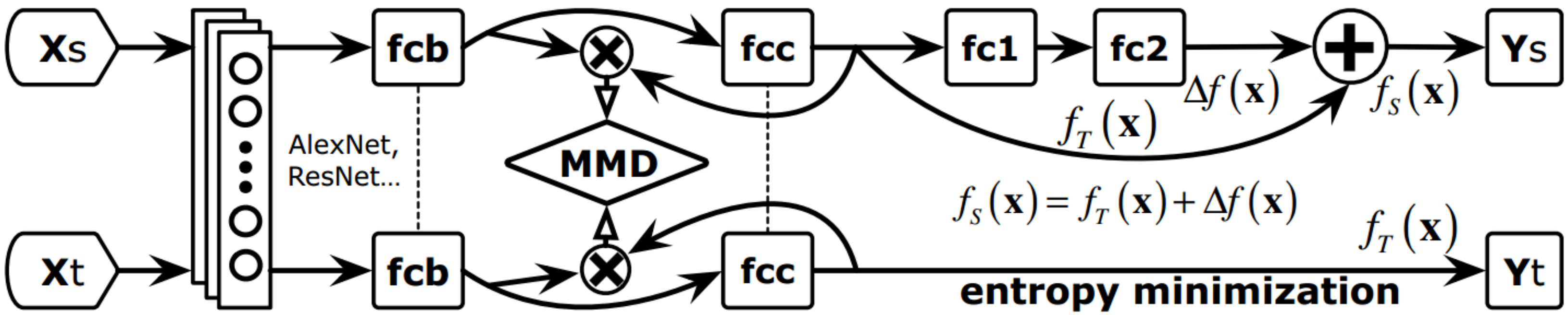
- Authors
- Motivation
- Methods
- Experiments

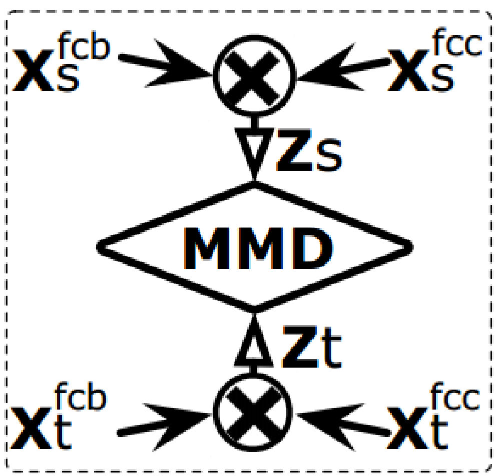
Main method

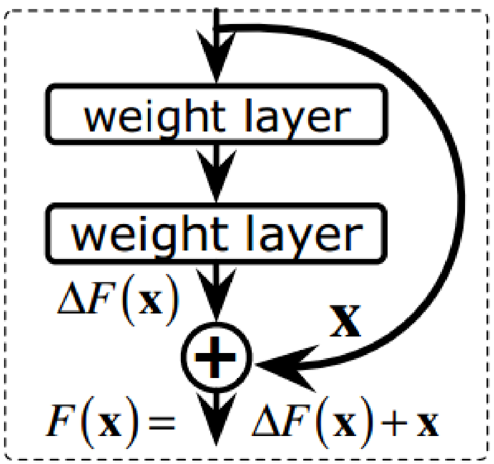
- High-level representation abstract: entropy objective + source domain regularizer

- Feature adaptation: joint training using Maximum Mean Discrepancy (MMD)

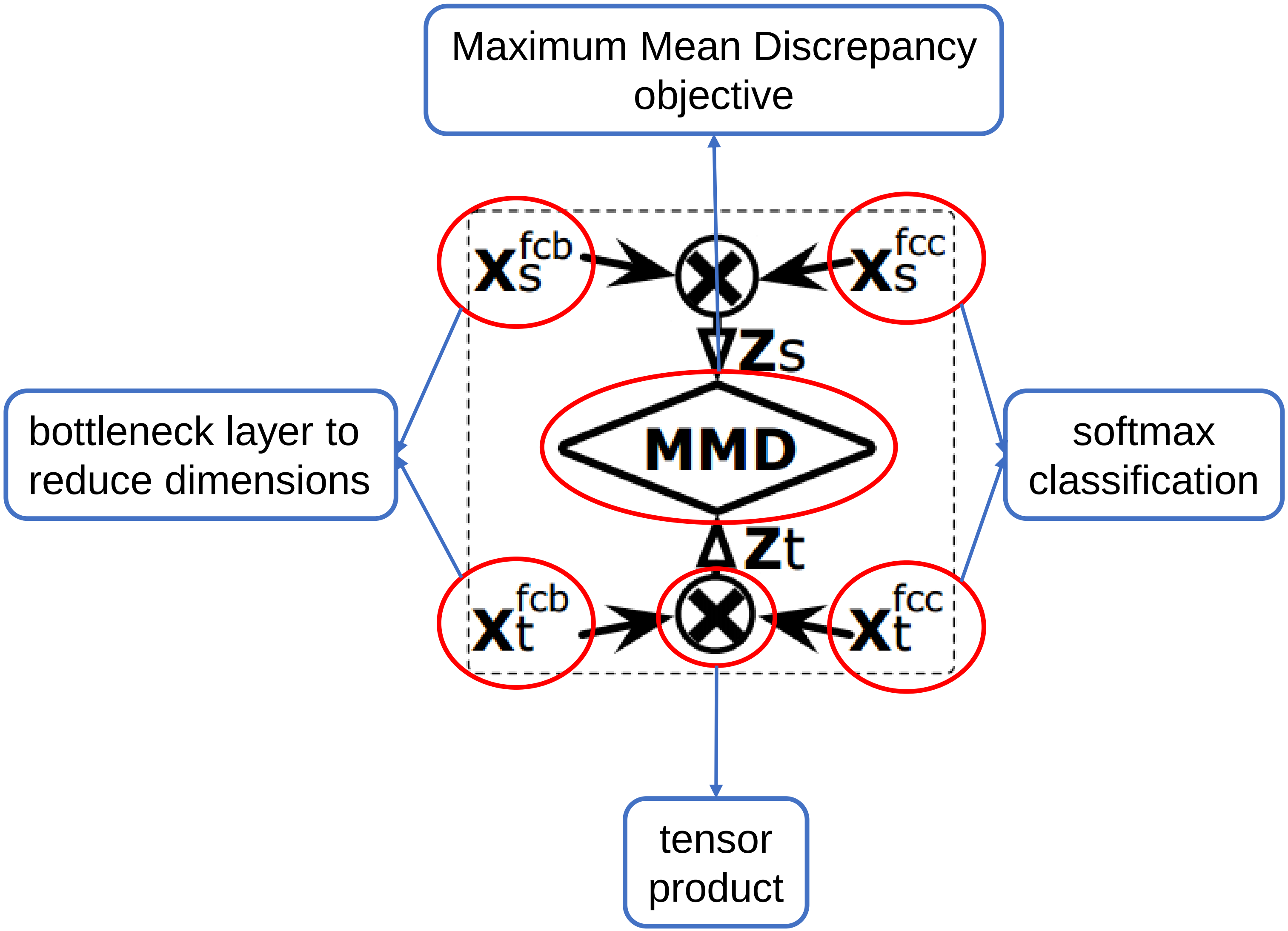
Architecture of Residual Transfer Network







Feature Adaptation



Maximum Mean Discrepancy

Maximum Mean Discrepancy (Fortet and Mourier, 1953)

$$D(p, q, \mathcal{F}) := \sup_{f \in \mathcal{F}} \mathbf{E}_p [f(x)] - \mathbf{E}_q [f(y)]$$

Theorem (via Dudley, 1984)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = C^0(\mathcal{X})$ is the space of continuous, bounded, functions on $\mathcal{X}$.

Theorem (via Steinwart, 2001; Smola et al., 2006)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = \{f | \, \|f\|_{\mathcal{H}} \leq 1\}$ is a unit ball in a Reproducing Kernel Hilbert Space, provided that $\mathcal{H}$ is universal.

Maximum Mean Discrepancy

Optimization Problem

$$\sup_{\|f\|\leq 1} \mathbf{E}_p\left[f(x)\right] - \mathbf{E}_q\left[f(y)\right] = \sup_{\|f\|\leq 1} \langle \mu_p - \mu_q, f\rangle = \|\mu_p - \mu_q\|_{\mathcal{H}}$$

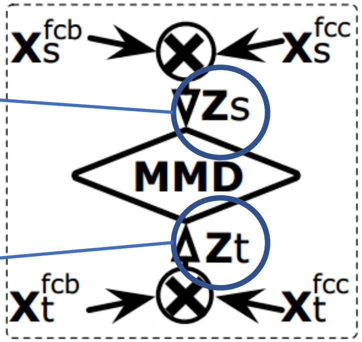
Kernels

$$\begin{aligned}
\|\mu_p - \mu_q\|_{\mathcal{H}}^2 &= \langle \mu_p - \mu_q, \mu_p - \mu_q\rangle \\
&= \mathbf{E}_{p,p}\left\langle k(x,\cdot), k(x',\cdot)\right\rangle - 2\mathbf{E}_{p,q}\left\langle k(x,\cdot), k(y,\cdot)\right\rangle \\
&\quad + \mathbf{E}_{q,q}\left\langle k(y,\cdot), k(y',\cdot)\right\rangle \\
&= \mathbf{E}_{p,p} k(x,x') - 2\mathbf{E}_{p,q} k(x,y) + \mathbf{E}_{q,q} k(y,y')
\end{aligned}$$

Maximum Mean Discrepancy

$$\mathbf{z}_i^s \triangleq \otimes_{\ell \in \mathcal{L}} \mathbf{X}_i^{s\ell}$$

$$\mathbf{z}_j^t \triangleq \otimes_{\ell \in \mathcal{L}} \mathbf{X}_j^{t\ell}.$$



$$k(\mathbf{z}, \mathbf{z}') = e^{-\left\|\mathrm{vec}(\mathbf{z}) - \mathrm{vec}(\mathbf{z}')\right\|^2 / b}$$

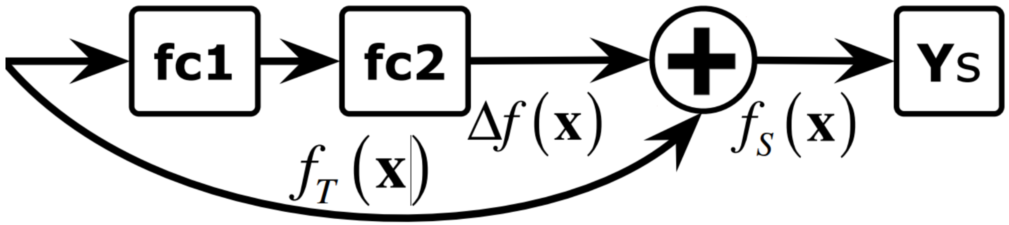
$$\min_{f_s, f_t} D_{\mathcal{L}}\left(\mathcal{D}_s, \mathcal{D}_t\right) = \sum_{i=1}^{n_s}\sum_{j=1}^{n_s} \frac{k\left(\mathbf{z}_i^s, \mathbf{z}_j^s\right)}{n_s^2} + \sum_{i=1}^{n_t}\sum_{j=1}^{n_t} \frac{k\left(\mathbf{z}_i^t, \mathbf{z}_j^t\right)}{n_t^2} - 2\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \frac{k\left(\mathbf{z}_i^s, \mathbf{z}_j^t\right)}{n_s n_t},$$

Classifier Adaptation

- Assume

$$f_S(\mathbf{x}) = f_T(\mathbf{x}) + \Delta f(\mathbf{x}),$$

- Residual connection



Classifier Adaptation

- Tackle unlabeled target domain data

$$\min_{f_t} \frac{1}{n_t} \sum_{i=1}^{n_t} H\left(f_t\left(\mathbf{x}_i^t\right)\right),$$

$$H\left(f_t\left(\mathbf{x}_i^t\right)\right) = - \sum_{j=1}^{c} f_j^t\left(\mathbf{x}_i^t\right) \log f_j^t\left(\mathbf{x}_i^t\right),$$

Bengio et al., semi supervised learning by entropy minimization

Jointly train：
Residual Transfer Network

$$\min_{f_S=f_T+\Delta f} \frac{1}{n_s} \sum_{i=1}^{n_s} L\left(f_s\left(\mathbf{x}_i^s\right), y_i^s\right) + \frac{\gamma}{n_t} \sum_{i=1}^{n_t} H\left(f_t\left(\mathbf{x}_i^t\right)\right) + \lambda\, D_{\mathcal{L}}\left(\mathcal{D}_s, \mathcal{D}_t\right),$$

Outline

- Authors
- Motivation
- Methods
- Experiments

Datasets

- **Office-31：4110 images in 31 classes from three domains**
  Amazon (**A**) *Webcam* (**W**) *DSLR* (**D**)

- **Office-Caltech**
  10 common categories shared by *Office-31* and *Caltech-256* (**C**), 12 transfer tasks

Results

Table 1: Accuracy on *Office-31* dataset using standard protocol [5] for unsupervised adaptation.

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
TCA [9]	59.0±0.0	90.2±0.0	88.2±0.0	57.8±0.0	51.6±0.0	47.9±0.0	65.8
GFK [14]	58.4±0.0	93.6±0.0	91.0±0.0	58.6±0.0	**52.4**±0.0	46.1±0.0	66.7
AlexNet [26]	60.6±0.4	95.4±0.2	99.0±0.1	64.2±0.3	45.5±0.5	48.3±0.5	68.8
DDC [4]	61.0±0.5	95.0±0.3	98.5±0.3	64.9±0.4	47.2±0.5	49.4±0.6	69.3
DAN [5]	68.5±0.3	96.0±0.1	99.0±0.1	66.8±0.2	50.0±0.4	49.8±0.3	71.7
RevGrad [6]	73.0±0.6	96.4±0.4	99.2±0.3	-	-	-	-
RTN (mmd)	70.0±0.4	96.1±0.3	99.2±0.3	67.6±0.4	49.8±0.4	50.0±0.3	72.1
RTN (mmd+ent)	71.2±0.3	96.4±0.2	99.2±0.1	69.8±0.2	50.2±0.3	50.7±0.2	72.9
RTN (mmd+ent+res)	**73.3**±0.3	**96.8**±0.2	**99.6**±0.1	**71.0**±0.2	50.5±0.3	**51.0**±0.1	**73.7**

Table 2: Accuracy on *Office-Caltech* dataset using standard protocol [5] for unsupervised adaptation.

Method	A→W	D→W	W→D	A→D	D→A	W→A	A→C	W→C	D→C	C→A	C→W	C→D	Avg
TCA [9]	84.4	96.9	99.4	82.8	90.4	85.6	81.2	75.5	79.6	92.1	88.1	87.9	87.0
GFK [14]	89.5	97.0	98.1	86.0	89.8	88.5	76.2	77.1	77.9	90.7	78.0	77.1	85.5
AlexNet [26]	79.5	97.7	**100.0**	87.4	87.1	83.8	83.0	73.0	79.0	91.9	83.7	87.1	86.1
DDC [4]	83.1	98.1	**100.0**	88.4	89.0	84.9	83.5	73.4	79.2	91.9	85.4	88.8	87.1
DAN [5]	91.8	98.5	**100.0**	91.7	90.0	92.1	84.1	81.2	80.3	92.0	90.6	89.3	90.1
RTN (mmd)	93.2	98.5	**100.0**	91.7	88.0	90.7	84.0	81.3	80.4	91.0	89.8	90.4	90.0
RTN (mmd+ent)	93.8	98.6	**100.0**	92.9	93.6	**92.7**	87.8	84.8	83.4	93.2	96.6	93.9	92.6
RTN (mmd+ent+res)	**95.2**	**99.2**	**100.0**	**95.5**	**93.8**	92.5	**88.1**	**86.6**	**84.6**	**93.7**	**96.9**	**94.2**	**93.4**

Thank you!