Auxiliary task guided mean and covariance alignment network for adversarial domain adaptation

Jiangmeng Li, Wenwen Qiang, Bing Su, Changwen Zheng
Knowledge-Based Systems (SCI Q1), 223 (2021) 107066
Jiangmeng Li, jiangmeng2019@iscas.ac.cn

TL;DR

Conventional adversarial domain adaptation (ADA) methods learn representations with strong transferability by eliminating the the Wasserstein distance-based discrepancy between the probability distributions of the source and the target domains and train the classifier only from the source domain data. We propose a novel method called auxiliary task guided mean and covariance alignment network (AT-MCAN) to take the second-order statistics differences into consideration and employ the data from both domains on training by introducing an auxiliary clustering task to the target domain.

Motivation

- The second-order statistic matters in transfer learning
- Exploring the unlabeled data in the target domain is valuable

Contribution

- We propose a new discrepancy metric be tween distributions that incorporates both the first-order and second-order statistics.
- We inroduce an auxiliary clustering task for the target domain to enhance the learned representations' discriminability.
- We provide theoretical analysis on the generation bound of the proposed metric and prove that introducing the auxiliary clustering task can promote the alignment between the label distributions of the source and target domains.

Methodology

- The proposed metric based on mean and covariance.
  - For gaining the first-order statistics, we avails of the Warsserstein distance-based discrepancy between distributions:

$$W_p\left(P_r, P_g\right) = \sup_{\|\gamma\|_L \le 1} \underset{x\sim P_r}{E}\left[\gamma\left(x\right)\right] - \underset{x\sim P_g}{E}\left[\gamma\left(x\right)\right],$$

where $\gamma : X \to R$ is the 1-Lipschitz function and satisfies $\|\gamma\|_L = \sup_{x\neq y}|\gamma\left(x\right) - \gamma\left(y\right)|/|x-y| \le 1$.

  - To take the second-order statistics into consideration, based on k orthogonal projection directions, we define a distance to maximize the discrimination between two domains:

$$D_{\mathrm{cov}} = \underset{x\sim P_r}{E}\left\|U^T\varphi\left(x\right)\left(V^T\varphi\left(x\right)\right)^T\right\|_a - \underset{y\sim P_g}{E}\left\|U^T\varphi\left(x\right)\left(V^T\varphi\left(x\right)\right)^T\right\|_a$$

where U, V are orthogonal matrices that satisfy $U^TU = I, V^TV = I$, I is the identity matrix of appropriate dimensions, $\varphi : X \to R$ is a 1-Lipschitz function which satisfies $\|\varphi\|_L = \sup_{x\neq y}|\varphi\left(x\right) - \varphi\left(y\right)|/|x-y| \le 1$, $\|.\|_a$ represents a certain matrix norm that can be nuclear norm, 1-norm, 2-norm and Frobenius norm and $P_r, P_g \in \mathrm{Prob}\left(X\right)$.

- The auxiliary clustering task guided classifier.

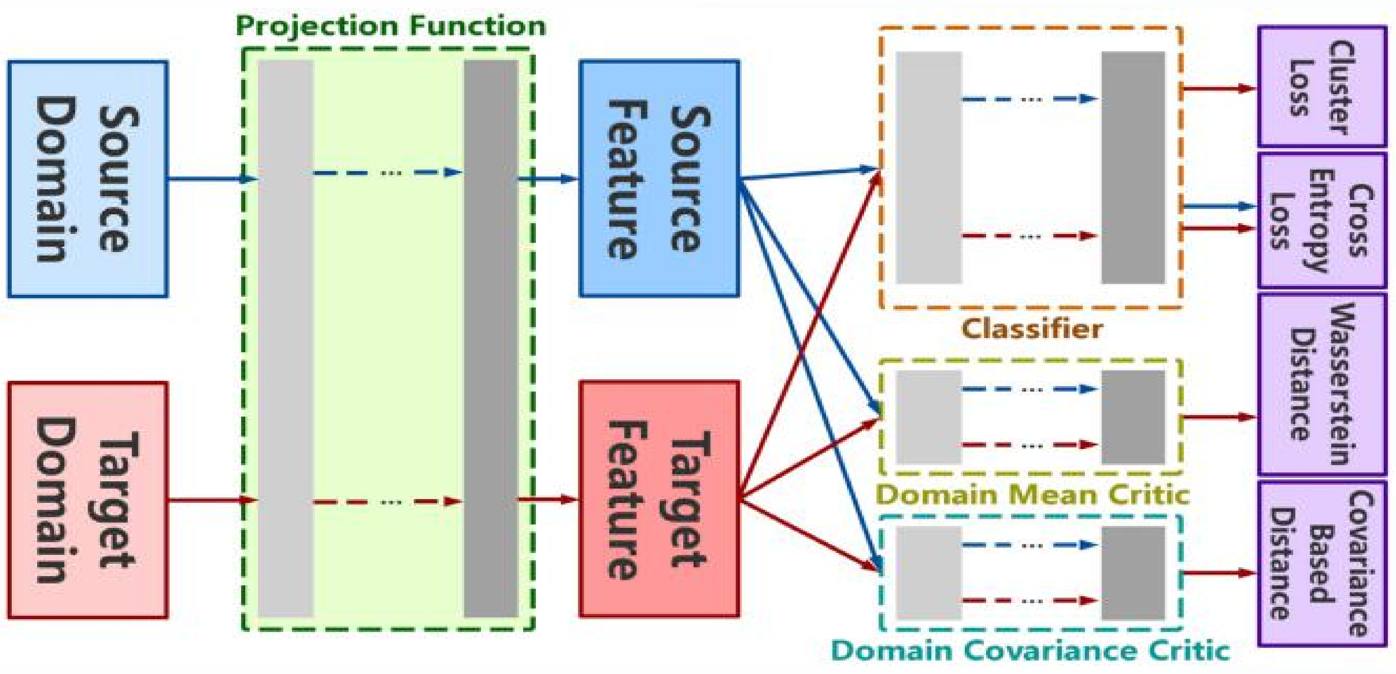


Fig. 1. The framework of AT-MCAN. First, AT-MCAN maps the data of the two domains to the latent space. Then, based on the feature representation of the latent space, AT-MCAN learns the classifier by minimizing clustering loss and cross-entropy loss, and aligns the distribution of the two domains by minimizing the proposed metric.

- We denote the clustering function by ς_{clu}, and the objective function is defined as:

$$R_{P_t^{clu}} = \frac{1}{n_t}\sum_{n=1}^{n_t} H\left(\varsigma_{clu}\left(f\left(x_n^t\right)\right)\right) - H\left(\frac{1}{n_t}\sum_{n=1}^{n_t}\varsigma_{clu}\left(f\left(x_n^t\right)\right)\right),$$

where n_t is the total number of samples in the target domain and H is an entropy operation.

Comparisons

- Classification comparisons to clarify the effectiveness of our proposed AT-MCAN:

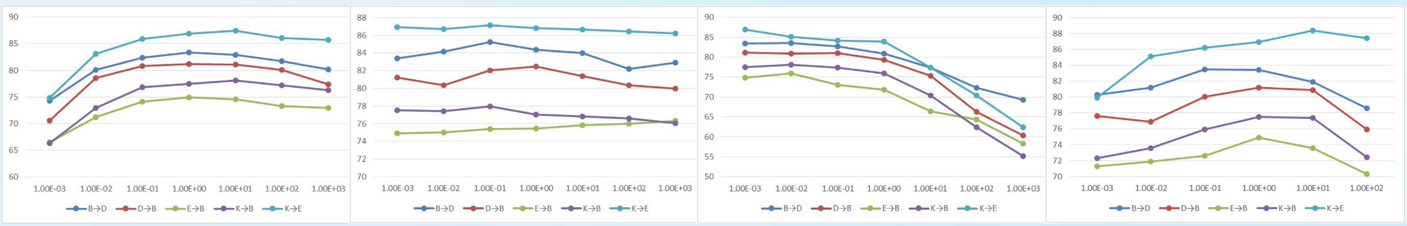
Performance (accuracy) on Office31 dataset.

Domains	A→D	A→W	D→A	D→W	W→A	W→D	Average
Source-only	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DAN	78.6	80.5	63.6	97.1	62.8	99.6	80.4
DANN	79.7	82.0	68.2	96.9	67.4	99.1	82.2
ADDA	77.8	86.2	69.5	96.2	68.9	98.4	82.8
JAN	84.7	85.4	68.6	97.4	70.0	99.8	84.3
MADA	87.8	90.0	70.3	97.4	66.4	99.6	85.3
SimNet	85.2	88.6	73.4	98.2	71.6	99.7	86.1
GTA	87.7	89.5	72.8	97.9	71.4	99.8	86.5
DAAA	88.8	86.8	74.3	99.3	73.9	**100.0**	87.2
CDAN	93.4	93.1	71.0	98.6	70.3	**100.0**	87.7
CAN	95.0	94.5	**78.0**	99.1	**77.0**	99.8	90.6
MEDA	86.2	85.9	72.3	97.4	73.4	99.4	85.8
SWD	83.5	82.5	75.7	88.9	72.5	96.4	83.3
CADA	**95.6**	97.0	71.5	99.3	73.1	**100.0**	89.5
MCAN	92.9	97.1	73.9	99.3	72.7	**100.0**	89.3
AT-MCAN	93.7	**97.7**	74.5	99.1	74.9	**100.0**	90.0
AT-MCAN*	**95.6**	**98.4**	76.2	**99.4**	76.5	**100.0**	**91.0**

Performance (accuracy) on Digits dataset.

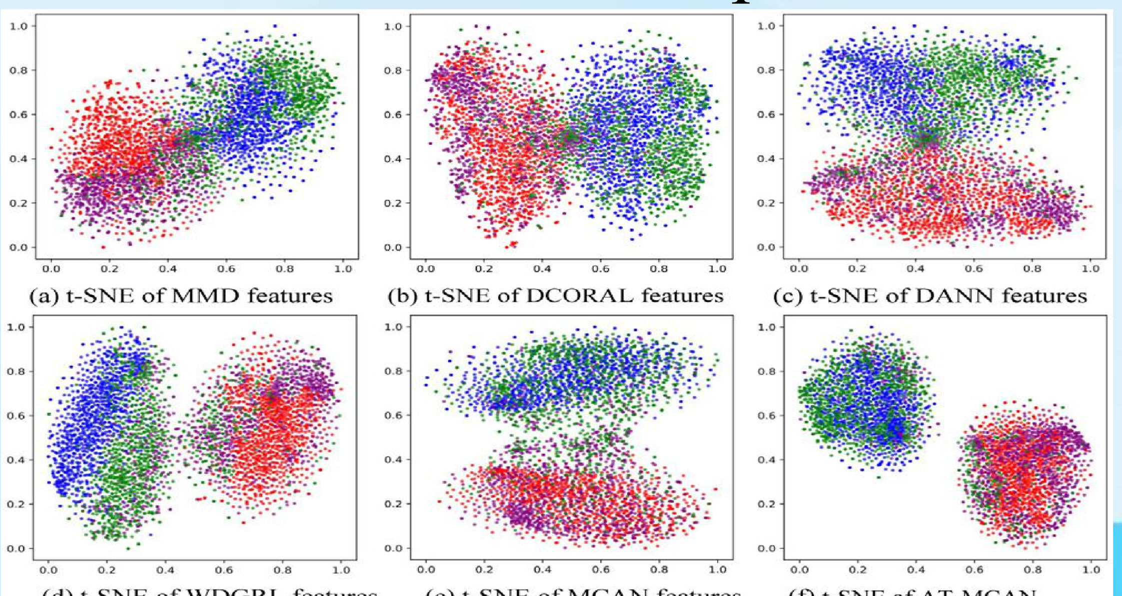
Domains	MNIST→USPS	USPS→MNIST	Average
DANN	90.4	94.7	92.6
ADDA	89.4	90.1	89.8
UNIT	**96.0**	93.6	94.8
CyCADA	95.6	96.5	96.1
CDAN	93.9	96.9	95.4
CDAN+E	95.6	98.0	96.8
BSP+DANN	94.5	97.7	96.1
BSP+ADDA	93.3	94.5	93.9
BSP+CDAN	95.0	**98.1**	96.6
SWD	**98.1**	97.1	97.6
MCAN	94.8	97.3	96.1
AT-MCAN	95.5	**98.1**	96.8
AT-MCAN*	97.6	**98.3**	**97.9**

- Studying the influences of the hyper-parameters, as follows:



(a) the influence of α (b) the influence of λ (c) the influence of θ (d) the influence of ι

- The visualization of the features learned by AT-MCAN and other compared methods:



(a) t-SNE of MMD features (b) t-SNE of DCORAL features (c) t-SNE of DANN features
(d) t-SNE of WDGRL features (e) t-SNE of MCAN features (f) t-SNE of AT-MCAN