Min-Max Statistical Alignment for Transfer Learning

Samitha Herath[1,4], Mehrtash Harandi[2,4], Basura Fernando[1,5], and Richard Nock[1,3,4]

[1]The Australian National University, [2]Monash University, [3]The University of Sydney, [4]DATA61-CSIRO, Australia

[5]Human-Centric AI Programme, A*STAR, Singapore

Samitha.Herath@data61.csiro.au, Mehrtash.Harandi@monash.edu, Fernando_Basura@scei.a-star.edu.sg, Richard.Nock@data61.csiro.au

Introduction

- We show that by minimizing the maximum (*i.e.*, min-max) statistical disparity, we can learn better domain invariant features.
- Our idea makes use of a novel structure, namely the *confusion network* to align distributions in a min-max framework.

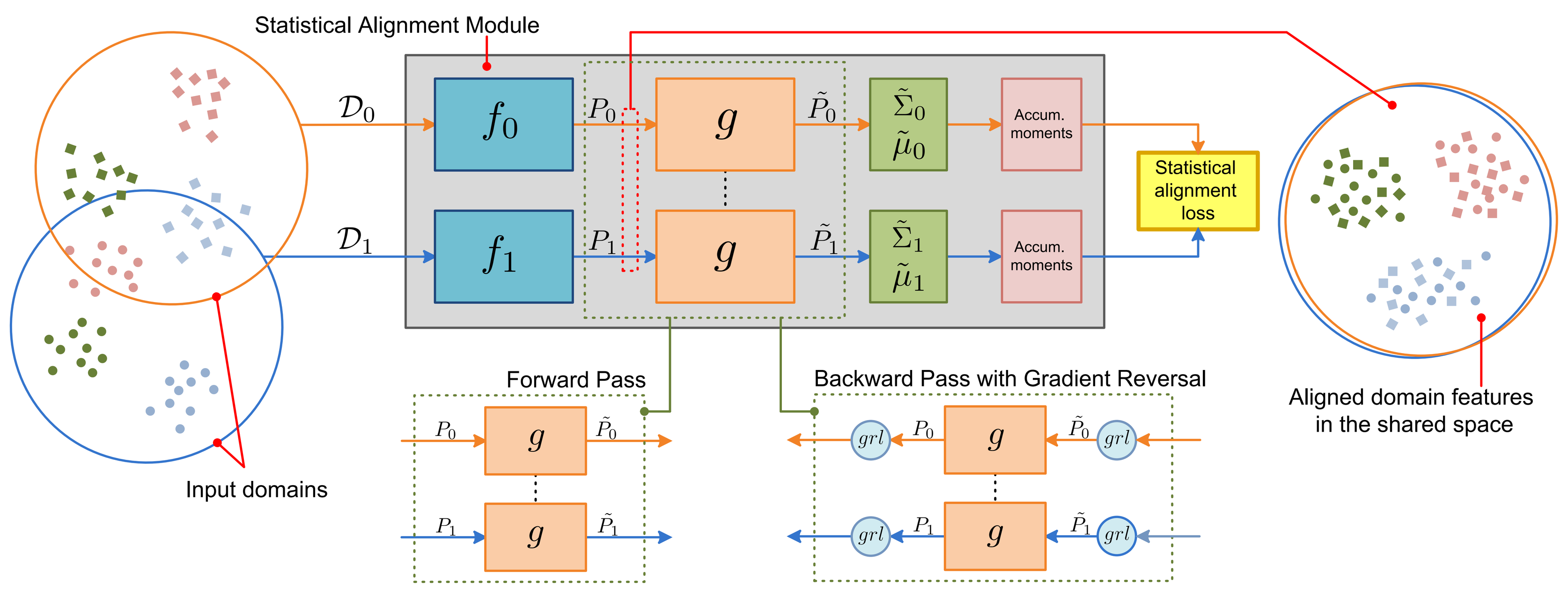

Figure: Schematic diagram for the proposed min-max statistical alignment. To realize the min-max training we propose using the confusion model, g.

Contributions

(i) We propose min-max statistical alignment using a novel confusion network.
(ii) We provide two frameworks to use min-max alignment for UDA and ZSL.

Statistical Alignment

Our objective is to learn two mappings, $f_k(\cdot, \boldsymbol{\theta}_k) : \mathbb{R}^{n_k} \to \mathbb{R}^d, k \in \{0, 1\}$ to embed samples from domains $\mathcal{D}_k, k \in \{0, 1\}$ to a shared feature space,

- Minimize a statistical disparity such as KL-divergence (D_{KL}) between domain feature distributions.
- Minimize the loss,

$$\mathcal{L}_u = \frac{1}{2}(D_{KL}(P_0 \| P_1) + D_{KL}(P_1 \| P_0)). \quad (1)$$

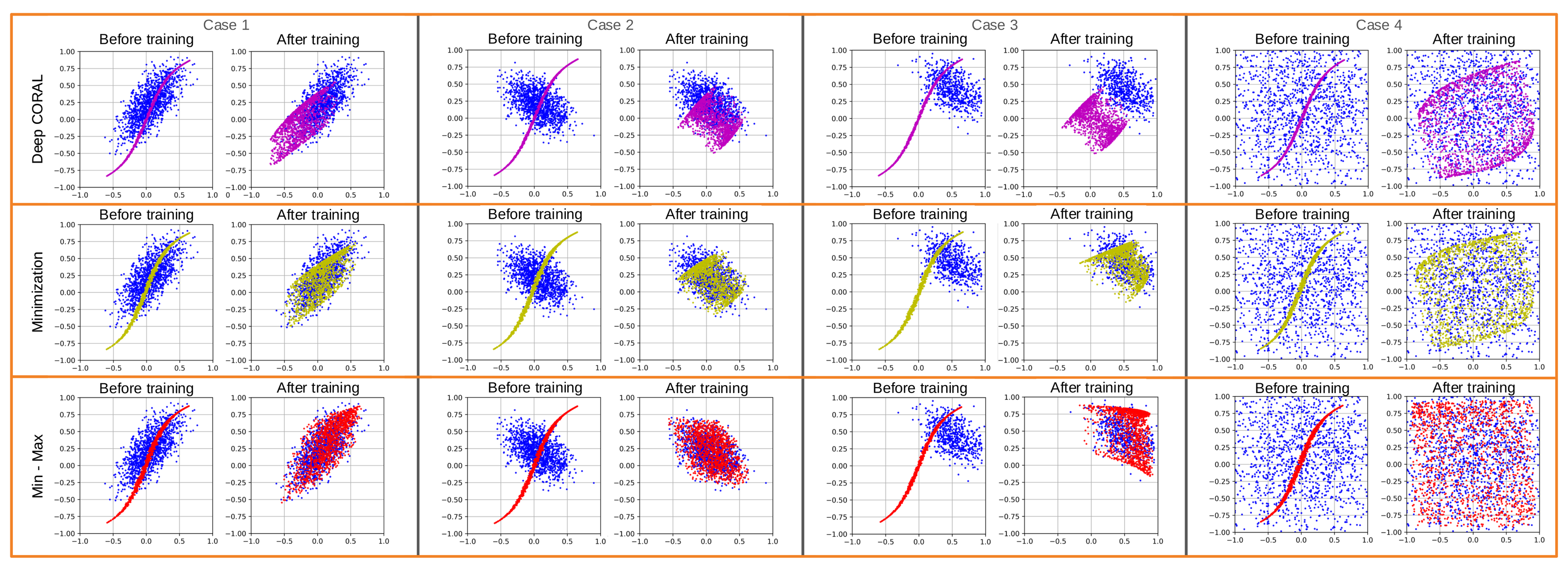

Figure: Toy-data experiment.

Min-Max Statistical Alignment

- We propose to make use of an additional mapping (*i.e.*, the *confusion network*), $g(\cdot, \boldsymbol{\theta}_g) : \mathbb{R}^d \to \mathbb{R}^p$.
- We perform min-max alignment by optimizing,

$$\min_{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1} \max_{\boldsymbol{\theta}_g} \tilde{\mathcal{L}}_u \ , \quad (2)$$

with,

$$\tilde{\mathcal{L}}_u = \frac{1}{2}(D_{KL}(\tilde{P}_0 \| \tilde{P}_1) + D_{KL}(\tilde{P}_1 \| \tilde{P}_0)). \quad (3)$$

Moment Accumulation

- When $\tilde{\boldsymbol{\Sigma}}^{(t)}$ and $\tilde{\boldsymbol{\mu}}^{(t)}$ are the computed covariance matrix and the mean vector for mini-batch at iteration t.
- When $0 \leq m < 1$, accumulated moments, $\tilde{\boldsymbol{\Sigma}}^{(t)}_{accu.}$ and $\tilde{\boldsymbol{\mu}}^{(t)}_{accu.}$ are computed as,

$$\tilde{\boldsymbol{\Sigma}}^{(t)}_{accu.} = m \times \tilde{\boldsymbol{\Sigma}}^{(t-1)}_{accu.} + (1 - m) \times \tilde{\boldsymbol{\Sigma}}^{(t)}, \quad (4)$$

$$\tilde{\boldsymbol{\mu}}^{(t)}_{\boldsymbol{accu.}} = m \times \tilde{\boldsymbol{\mu}}^{(t-1)}_{accu.} + (1 - m) \times \tilde{\boldsymbol{\mu}}^{(t)}. \quad (5)$$

Case study 1 : Unsupervised Domain Adaptation

We apply min-max alignment for UDA with $\mathcal{D}_0 \sim \mathcal{D}_s$ and $\mathcal{D}_1 \sim \mathcal{D}_t$.

$$\min_{\boldsymbol{\theta}_s, \boldsymbol{\theta}_t, \boldsymbol{\theta}_h} \max_{\boldsymbol{\theta}_g} \mathcal{L}_{d,s} + \lambda_t \mathcal{L}_{d,t} + \lambda_u \tilde{\mathcal{L}}_u. \quad (6)$$

Here, $\mathcal{L}_{d,s}$ is the softmax cross-entropy loss computed using the labeled source samples and $\mathcal{L}_{d,t}$ is the unsupervised discriminative loss,

$$\mathcal{L}_{d,t} = -\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_t}[h_t(\boldsymbol{x})^T \log h_t(\boldsymbol{x})].$$

Sol.	s	A	A	D	D	W	W	MNIST	SVHN	DIGITS	SIGNS	STL	CIFAR
	t	D	W	A	W	A	D	SVHN	MNIST	SVHN	GTSRB	CIFAR	STL
CNN		60.8	58.5	42.6	94.1	38.6	98.1	37.5	63.1	84.0	77.8	59.1	75.9
Min		66.1	70.5	47.2	93.5	47.6	98.6	49.8	70.6	85.4	78.7	57.7	74.4
Min-Max		**69.1**	**71.7**	**51.3**	**95.0**	**52.1**	**99.3**	**67.8**	**72.0**	**88.5**	**83.2**	**60.6**	**76.1**

Table: **Comparison of the Min-Max UDA solution with CNN and Min.** For Office31 we use the "fully transductive protocol" of [1,2]. For the remaining sets we follow [3].

Sol.	s	A	A	D	D	W	W	MNIST	SVHN	DIGITS	SIGNS	STL	CIFAR
	t	D	W	A	W	A	D	SVHN	MNIST	SVHN	GTSRB	CIFAR	STL
DANN		67.1	**73.0**	**54.5**	**96.4**	**52.7**	99.2	60.6	68.3	90.1	97.5	62.7	78.1
VADA		-	-	-	-	-	-	73.3	94.5	**94.9**	**99.2**	**71.4**	78.3
D-CORAL		66.8	66.4	52.8	95.7	51.5	99.2	72.7	87.8	71.8	59.9	60.5	76.2
Min-Max+		**69.1**	71.7	51.3	95.0	52.1	**99.3**	**79.3**	**97.0**	94.6	97.3	67.7	**79.9**

Table: **Comparison of the Min-Max+ UDA solution with related solutions.** For Office31 we use the "fully transductive protocol" of [1,2]. For the remaining sets we follow [3].

Case study 2 : Zero-Shot Learning

We apply min-max alignment for ZSL with $\mathcal{D}_0 \sim \mathcal{D}_s$ and $\mathcal{D}_1 \sim \mathcal{D}_{att.}$

$$\min_{\boldsymbol{\theta}_s, \boldsymbol{\theta}_h} \mathcal{L}_{d,s}, \quad (7)$$

$$\min_{\boldsymbol{\theta}_{att.}} \max_{\boldsymbol{\theta}_g} \mathcal{L}_{d,att.} + \lambda_u \tilde{\mathcal{L}}_u. \quad (8)$$

Data.	Awa1			Awa2			Cubs			Sun		
Sol.	U	S	HM	U	S	HM	U	S	HM	U	S	HM
SAE	1.8	77.1	3.5	1.1	82.2	2.2	7.8	54.0	13.6	8.8	18.0	11.8
ZKL	18.3	79.3	29.7	18.9	82.7	30.8	24.2	63.9	35.1	21.0	31.0	25.1
Cls. Prot.	28.1	73.5	40.6	-	-	-	23.5	55.2	32.9	21.5	34.7	26.5
CLSW	57.9	61.4	59.6	-	-	-	43.7	57.7	49.7	42.6	36.6	**39.4**
Min	46.0	83.3	59.3	32.9	89.7	48.1	46.1	50.8	48.3	38.8	35.0	36.8
Min-Max	46.6	84.2	**60.0**	37.8	88.8	**53.0**	47.1	53.8	**50.2**	37.9	36.5	37.2

Table: Comparison of the proposed ZSL solution (Min-Max) on GZSL protocol [4].

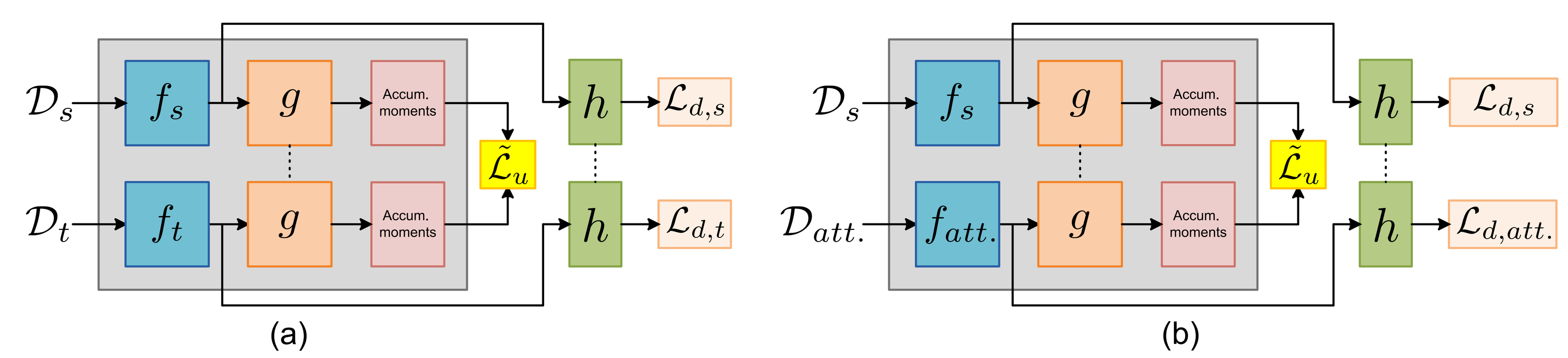

Figure: Schematic diagram for (a) UDA and (b) ZSL models with the proposed alignment.

Further Study : Image Generation

We further use our min-max alignment to generate images using DC-GAN network models.

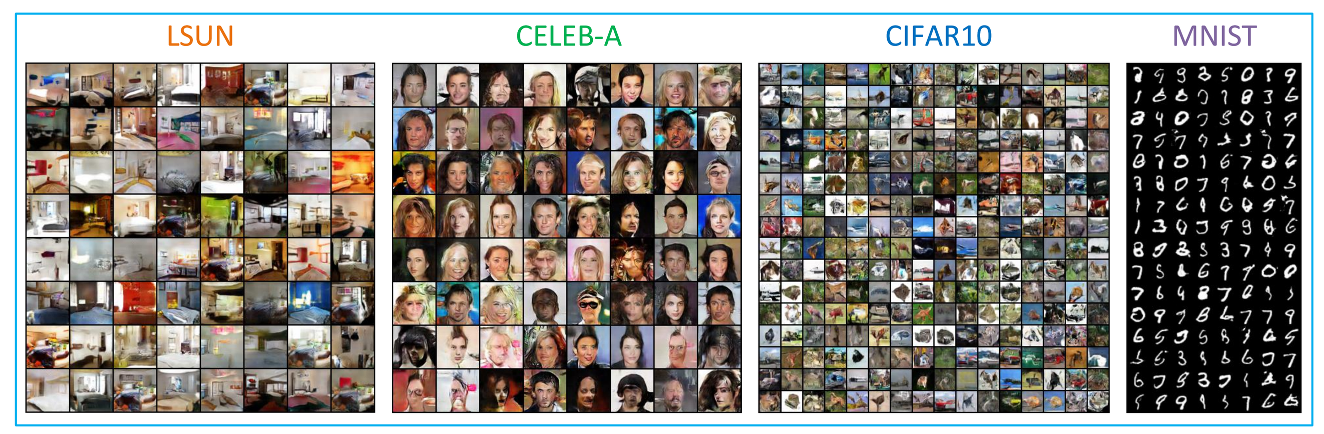

Figure: Generated images using min-max alignment with DC-GAN networks.

Solution	IS
Real Img	11.95
W-GAN	5.88
Imp. GAN	4.36
Mmd GAN	**6.17**
Min-Max	5.92

Table: Inception Scores (IS) [5] for CIFAR10.

References.

[1] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1):20962030, 2016.

[2] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In Domain Adaptation in Computer Vision Applications, pages 153171. Springer, 2017.

[3] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In International Conference on Learning Representations, 2018.

[4] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 30773086, 2017.

[5] Li, C.L., Chang, W.C., Cheng, Y., Yang, Y. and Pczos, B., 2017. Mmd gan: Towards deeper understanding of moment matching network. In Advances in Neural Information Processing Systems (pp. 2203-2213).