Information-Theoretic Analysis of Unsupervised Domain Adaptation

Ziqiao Wang[1] Yongyi Mao[1]

[1]School of EECS, University of Ottawa

Outline

Background

- Unsupervised Domain Adaptation (UDA): leveraging both labeled source domain data and unlabeled target domain data to carry out various tasks in the target domain

Background

- Unsupervised Domain Adaptation (UDA): leveraging both labeled source domain data and unlabeled target domain data to carry out various tasks in the target domain
- *Nguyen, A. Tuan, et al. "KL Guided Domain Adaptation." ICLR 2022.*

Background

- Unsupervised Domain Adaptation (UDA): leveraging both labeled source domain data and unlabeled target domain data to carry out various tasks in the target domain
- *Nguyen, A. Tuan, et al. "KL Guided Domain Adaptation." ICLR 2022.*
  - Representation Network:
    - Input: data
    - Output: a mean vector $\hat{\mu} \in \mathbb{R}^d$ and a variance vector $\hat{\sigma}^2 \in \mathbb{R}^d$
    - Gaussian of source domain $\mathcal{N}(\hat{\mu}_{\mathrm{s}}, \hat{\sigma}^2_{\mathrm{s}}\mathrm{I}_d)$; Gaussian of target domain $\mathcal{N}(\hat{\mu}_{\mathrm{t}}, \hat{\sigma}^2_{\mathrm{t}}\mathrm{I}_d\})$
    - Minimizing KL divergence between two Gaussian distributions

Background

- Unsupervised Domain Adaptation (UDA): leveraging both labeled source domain data and unlabeled target domain data to carry out various tasks in the target domain
- *Nguyen, A. Tuan, et al. "KL Guided Domain Adaptation." ICLR 2022.*
  - Representation Network:
    - Input: data
    - Output: a mean vector $\hat{\mu} \in \mathbb{R}^d$ and a variance vector $\hat{\sigma}^2 \in \mathbb{R}^d$
    - Gaussian of source domain $\mathcal{N}(\hat{\mu}_{\mathrm{s}}, \hat{\sigma}_{\mathrm{s}}^2 \mathrm{I}_d)$; Gaussian of target domain $\mathcal{N}(\hat{\mu}_{\mathrm{t}}, \hat{\sigma}_{\mathrm{t}}^2 \mathrm{I}_d\})$
    - Minimizing KL divergence between two Gaussian distributions
  - Classifier:
    - Sampling from $\mathcal{N}(\hat{\mu}_{\mathrm{s}}, \hat{\sigma}_{\mathrm{s}}^2 \mathrm{I}_d)$
    - Minimizing cross-entropy loss

Outline

Notations

- Instance space: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Hypothesis space: $\mathcal{W} \subseteq \mathbb{R}^d$; Predictor space: $\mathcal{F} = \{f_w : \mathcal{X} \to \mathcal{Y} | w \in \mathcal{W}\}$
- Source data $Z = (X, Y) \sim \mu$ and target data $Z' = (X', Y') \sim \mu'$
- Labeled source sample: $S = \{Z_i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mu^{\otimes n}$; Unlabelled target sample $S'_{X'} = \{X'_j\}_{j=1}^m \overset{\text{i.i.d}}{\sim} P_{X'}^{\otimes m}$
- Learning algorithm: $\mathcal{A} : \mathcal{Z}^n \times \mathcal{X}^m \to \mathcal{W}$ by $P_{W|S,S'_{X'}}$

Generalization Error

- Loss: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$
- We're interested in
  - Population risk of target domain: $R_{\mu'}(w) \triangleq \mathbb{E}_{Z'}[\ell(f_w(X'), Y')]$
  - Empirical risk of source domain: $R_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(X_i), Y_i)$
  - *Expected empirical-to-population (EP) generalization error*:

$$\mathrm{Err} \triangleq \mathbb{E}_{W,S}\left[R_{\mu'}(W) - R_S(W)\right] = \mathbb{E}_{W,S,S'_{X'}}\left[R_{\mu'}(W) - R_S(W)\right]$$

  - *Population-to-population (PP) generalization error for* w:

$$\widetilde{\mathrm{Err}}(w) \triangleq R_{\mu'}(w) - R_{\mu}(w)$$

  - Relation between EP and PP:

$$|R_{\mu'}(w) - R_S(w)| \leq |R_{\mu'}(w) - R_{\mu}(w)| + |R_{\mu}(w) - R_S(w)|$$

Assumptions of the Loss Function

1 Boundedness: $\ell(\cdot,\cdot)$ is bounded in $[0, M]$.

2 Subgaussianity: $\ell(f_w(X), Y)$ is R-subgaussian [1] under μ for any $w \in \mathcal{W}$.

3 Lipschitzness: $\ell(f_w(X), Y)$ is β-Lipschitz continuous in $\mathcal{Z}$ w.r.t. a metric d for any $w \in \mathcal{W}$, i.e., $|\ell(f_w(x_1), y_1) - \ell(f_w(x_2), y_2)| \leq \beta d(z_1, z_2)$.

4 Triangle and Symmetric: $\ell(\cdot,\cdot)$ satisfies the following: $\ell(y_1, y_2) = \ell(y_2, y_1)$ and $\ell(y_1, y_2) \leq \ell(y_1, y_3) + \ell(y_3, y_2)$ *for any* $y_1, y_2, y_3 \in \mathcal{Y}$.

[1] A random variable X is R-subgaussian if for any ρ, $\log \mathbb{E} \exp(\rho(X - \mathbb{E}X)) \leq \rho^2 R^2/2$.

Outline

Main Ingredients

Lemma 1 (Donsker and Varadhan's variational formula)

Let Q, P be probability measures on Θ, for any bounded measurable function $f : \Theta \to \mathbb{R}$, we have $\mathrm{D_{KL}}(Q||P) = \sup_f \mathbb{E}_{\theta\sim Q}\left[f(\theta)\right] - \log \mathbb{E}_{\theta\sim P}\left[\exp f(\theta)\right]$.

Also called "change of measure inequality", " the Legendre transform of KL divergence" ...

Lemma 2

Let Q and P be probability measures on Θ. Let $\theta' \sim Q$ and $\theta \sim P$. If $g(\theta)$ is R-subgaussian, then,

$$\left|\mathbb{E}_{\theta'\sim Q}\left[g(\theta')\right] - \mathbb{E}_{\theta\sim P}\left[g(\theta)\right]\right| \leq \sqrt{2R^2\mathrm{D_{KL}}(Q||P)}.$$

Bounding PP Error by KL Divergence

Theorem 1

If Assumption 2 holds, then for any $w \in \mathcal{W}$, $\left|\widetilde{\mathrm{Err}}(w)\right| \leq \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}(\mu' || \mu)}$.

Bounding PP Error by KL Divergence

Theorem 1

If Assumption 2 holds, then for any $w \in \mathcal{W}$, $\left|\widetilde{\mathrm{Err}}(w)\right| \leq \sqrt{2R^2 \mathrm{D_{KL}}(\mu' || \mu)}$.

Corollary 1

Suppose that $f_w = g \circ h$ *(where* h *is a function mapping* $\mathcal{X}$ *to a representation space* $\mathcal{T}$ *and* g *is a function mapping* $\mathcal{T}$ *to* $\mathcal{Y}$*) and that Assumption 2 holds. then for any* $w \in \mathcal{W}$,

$$R_\mu(w) - \sqrt{2R^2 \mathrm{D_{KL}}(\mu' || \mu)} \leq R_{\mu'}(w) \leq R_\mu(w) + \sqrt{2R^2 \mathrm{D_{KL}}(\mu'_{\mathrm{h}} || \mu_{\mathrm{h}})}.$$

Bounding PP Error by KL Divergence

Theorem 1

If Assumption 2 holds, then for any $w \in \mathcal{W}$, $\left|\widetilde{\mathrm{Err}}(w)\right| \leq \sqrt{2R^2 \mathrm{D_{KL}}(\mu' || \mu)}$.

Corollary 1

Suppose that $f_w = g \circ h$ *(where* h *is a function mapping* $\mathcal{X}$ *to a representation space* $\mathcal{T}$ *and* g *is a function mapping* $\mathcal{T}$ *to* $\mathcal{Y}$*) and that Assumption 2 holds. then for any* $w \in \mathcal{W}$,

$$R_\mu(w) - \sqrt{2R^2 \mathrm{D_{KL}}(\mu' || \mu)} \leq R_{\mu'}(w) \leq R_\mu(w) + \sqrt{2R^2 \mathrm{D_{KL}}(\mu'_{\mathrm{h}} || \mu_{\mathrm{h}})}.$$

Corollary 2

If Assumption 1 holds,

$$\left|\widetilde{\mathrm{Err}}(w)\right| \leq \frac{M}{\sqrt{2}} \sqrt{\min\{\mathrm{D_{KL}}(\mu || \mu'), \mathrm{D_{KL}}(\mu' || \mu)\}} \leq \frac{M}{2} \sqrt{\mathrm{D_{KL}}(\mu || \mu') + \mathrm{D_{KL}}(\mu' || \mu)}.$$

Incorrect pseudo labels may even hurt the target domain performance.

Suppose the trained model Q can well approximate the real mapping between X and Y on source domain (i.e. $Q_{Y|T} = P_{Y|T}$).

Let $\hat{Y}'$ be the pseudo label of T' generated by the trained model, i.e., $Q_{\hat{Y}'|T'} = Q_{Y|T}$. Let $Q_{T',\hat{Y}'} = P_{T'}Q_{\hat{Y}'|T'}$, then

$$\mathrm{D_{KL}}(P_{T',Y'}||P_{T,Y}) = \mathbb{E}_{P_{T',Y'}} \log \frac{P_{T',Y'}Q_{T',\hat{Y}'}}{Q_{T',\hat{Y}'}P_{T,Y}} = \mathrm{D_{KL}}(P_{T'}||P_T) + \mathrm{D_{KL}}(P_{Y'|T'}||Q_{\hat{Y}'|T'}). \tag{1}$$

For a specific t' and y', if $P(Y' = y'|T' = t') \neq 0$ and $Q(\hat{Y}' = y'|T' = t') = 0$, then the second term in RHS of Eq. (1), $\mathrm{D_{KL}}(P_{Y'|T'}||Q_{\hat{Y}'|T'}) \to \infty$.

Theorem 2

If Assumption 4 holds and let $\ell(f_{w'}(X), f_w(X))$ be R-subgaussian for any $w, w' \in \mathcal{W}$. Then for any w, $\widetilde{\mathrm{Err}}(w) \leq \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}(P_{X'}||P_X)} + \lambda^*$, where $\lambda^* = \min_{w \in \mathcal{W}} R_{\mu'}(w) + R_\mu(w)$.

Here λ^* measures the possibility of whether the domain adaptation algorithm will succeed under the oracle knowledge of μ and μ'.

Bounding PP Error by Wasserstein Distance

Theorem 3

If Assumption 3 holds, then $\left|\widetilde{\mathrm{Err}}(w)\right| \leq \beta \mathbb{W}(\mu', \mu)$.

Main tool: Kantorovich–Rubinstein duality of Wasserstein distance

Lemma 3 (KR duality)

For any two distributions P and Q, we have

$$\mathbb{W}(P, Q) = \sup_{f \in 1-\mathrm{Lip}(\rho)} \int_{\mathcal{X}} f dP - \int_{\mathcal{X}} f dQ,$$

where the supremum is taken over all 1-Lipschitz functions in the metric d, i.e. $|f(x) - f(x')| \leq d(x, x')$ for any $x, x' \in \mathcal{X}$.

Corollary 3

If Assumption 1 holds and let d be the discrete metric, then

$$\left|\widetilde{\mathrm{Err}}(w)\right| \leq M\mathrm{TV}(\mu', \mu) \leq M\sqrt{\min\left\{\frac{1}{2}\mathrm{D_{KL}}(\mu'||\mu), 1 - e^{-\mathrm{D_{KL}}(\mu'||\mu)}\right\}}.$$

Main tool for the second inequality: Pinsker's inequality and Bretagnolle-Huber inequality.

Theorem 4

If Assumption 4 holds and $\ell(f_w(X), f_{w'}(X))$ is β-Lipschitz in $\mathcal{X}$ for any $w, w' \in \mathcal{W}$, then for any $w \in \mathcal{W}$, $\widetilde{\mathrm{Err}}(w) \leq \beta\mathbb{W}(P_{X'}, P_X) + \lambda^*$, where $\lambda^* = \min_{w\in\mathcal{W}} R_{\mu'}(w) + R_{\mu}(w)$.

Outline

Additional Prerequisites

Definition 1 (Disintegrated Mutual Information)

Let X, Y and Z be random variables and z be a realization of Z. The disintegrated mutual information of X and Y given $Z = z$ is $I^z(X;Y) \triangleq \mathrm{D}_{\mathrm{KL}}(P_{X,Y|Z=z}||P_{X|Z=z}P_{Y|Z=z})$.

Note that the conditional mutual information $I(X;Y|Z) = \mathbb{E}_Z I^Z(X;Y)$.

Definition 2 (Lautum Information)

The lautum information between X and Y is defined as

$$L(X;Y) \triangleq \mathrm{D}_{\mathrm{KL}}(P_X P_Y||P_{XY}).$$

MI Bound for EP

Theorem 5

Assume $\ell(f_w(X'), Y')$ is R-subgaussian under μ' for any $w \in \mathcal{W}$. Then

$$|\mathrm{Err}| \leq \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2 I^{X'_j}(W;Z_i)} + \sqrt{2R^2 \mathrm{D_{KL}}(\mu||\mu')}.$$

MI Bound for EP

Theorem 5

Assume $\ell(f_w(X'), Y')$ is R-subgaussian under μ' for any $w \in \mathcal{W}$. Then

$$|\mathrm{Err}| \leq \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2 I^{X'_j}(W;Z_i)} + \sqrt{2R^2 \mathrm{D_{KL}}(\mu||\mu')}.$$

Remark 1

Moving the expectation inside the square root function by Jensen's ineq. By $Z_i \perp\!\!\!\perp X'_j$, we have

$$I(W;Z_i|X'_j) = I(W;Z_i|X'_j) + I(Z_i;X'_j) = I(W;Z_i) + I(X'_j;Z_i|W).$$

The term $I(W;Z_i)$ will vanish as $n \to \infty$ and the term $I(X'_j;Z_i|W)$ will also vanish as $n, m \to \infty$.

Stronger Bounds

Corollary 4

Let Assumption 1 hold. Then

$$|\mathrm{Err}| \leq \frac{M}{\sqrt{2}nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{\min\left\{I^{X'_j}(W;Z_i), L^{X'_j}(W;Z_i)\right\}} + \frac{M}{\sqrt{2}}\sqrt{\min\left\{\mathrm{D_{KL}}(\mu||\mu'), \mathrm{D_{KL}}(\mu'||\mu)\right\}},$$

where $L^{X'_j}(\cdot;\cdot)$ *is the disintegrated version of Lautum information.*

Stronger Bounds

Theorem 6

Assume ℓ is Lipschitz for both $w \in \mathcal{W}$ and $z \in \mathcal{Z}$, i.e., $|\ell(f_w(x),y) - \ell(f_w(x'),y')| \leq \beta d_1(z,z')$ for all $z, z' \in \mathcal{Z}$ and $|\ell(f_w(x),y) - \ell(f_{w'}(x),y)| \leq \beta' d_2(w,w')$ for all $w, w' \in \mathcal{W}$, then

$$|\mathrm{Err}| \leq \frac{\beta'}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{E}_{X'_j, Z_i} \mathbb{W}(P_{W|Z_i, X'_j}, P_{W|X'_j}) + \beta \mathbb{W}(\mu, \mu').$$

Further, if Assumption 1 hold. Then

$$\left|\widetilde{\mathrm{Err}}\right| \leq \frac{M}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{E}_{X'_j, Z_i} \left[\mathrm{TV}(P_{W|Z_i, X'_j}, P_{W|X'_j})\right] + M\mathrm{TV}(\mu, \mu')$$
$$\leq \frac{1}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{E}_{X'_j, Z_i} \sqrt{\frac{M^2}{2} \mathrm{D_{KL}}(P_{W|Z_i, X'_j} || P_{W|X'_j})} + \sqrt{\frac{M^2}{2} \mathrm{D_{KL}}(\mu || \mu')}.$$

Outline

Application: Gradient Penalty as an Universal Regularizer

Consider a “noisy” iterative algorithm for updating W, e.g., SGLD. At each time step t,

- labelled source mini-batch: Z_{B_t}
- unlabelled target mini-batch: X'_{B_t}
- gradient: $g(W_{t-1}, Z_{B_t}, X'_{B_t})$
- updating rule: $W_t = W_{t-1} - \eta_t g(W_{t-1}, Z_{B_t}, X'_{B_t}) + N_t$ where η_t is the learning rate and $N_t \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$.

Application: Gradient Penalty as an Universal Regularizer

Theorem 7

Under the assumption of Theorem 5. Let the total iteration number be T and let $G_t = g(W_{t-1}, Z_{B_t}, X'_{B_t})$, then

$$|\mathrm{Err}| \leq \sqrt{\frac{R^2}{n}\sum_{t=1}^{T}\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{S'_{X'},W_{t-1},S}\left[\left|\left|G_t - \mathbb{E}_{Z_{B_t}}\left[G_t\right]\right|\right|^2\right]} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}(\mu||\mu')}.$$

restrict the gradient norm $\Longrightarrow$ reduce $|\mathrm{Err}|$.

This strategy will also restrict the distance between the final output W_T and the initialization W_0, effectively shrinking the hypothesis space accessible by the algorithm.

Application: Controlling Label Information for KL Guided Marginal Alignment

Motivation: Discrepancy between $P_{Y|T}$ and $P_{Y'|T'}$

- Nguyen et al. (2022) shows that $\mathrm{D_{KL}}(P_{Y'|T'}||P_{Y|T}) \leq \mathrm{D_{KL}}(P_{Y'|X'}||P_{Y|X})$ if $I(X;Y) = I(T;Y)$. Penalizing the KL of the marginals is safe.

Application: Controlling Label Information for KL Guided Marginal Alignment

Motivation: Discrepancy between $P_{Y|T}$ and $P_{Y'|T'}$

- Nguyen et al. (2022) shows that $\mathrm{D_{KL}}(P_{Y'|T'}||P_{Y|T}) \leq \mathrm{D_{KL}}(P_{Y'|X'}||P_{Y|X})$ if $I(X;Y) = I(T;Y)$. Penalizing the KL of the marginals is safe.
- The condition $I(X;Y) = I(T;Y)$ can be difficult to satisfy when ℓ is cross-entropy.

Application: Controlling Label Information for KL Guided Marginal Alignment

Motivation: Discrepancy between $P_{Y|T}$ and $P_{Y'|T'}$

- Nguyen et al. (2022) shows that $\mathrm{D_{KL}}(P_{Y'|T'}||P_{Y|T}) \leq \mathrm{D_{KL}}(P_{Y'|X'}||P_{Y|X})$ if $I(X;Y) = I(T;Y)$. Penalizing the KL of the marginals is safe.
- The condition $I(X;Y) = I(T;Y)$ can be difficult to satisfy when ℓ is cross-entropy.
  By DPI on $Y - X - T$, $I(X;Y) \geq I(T;Y) = H(Y) - H(Y|T)$.

$$\mathbb{E}_{W,Z_i}\left[\ell(f_W(T_i),Y_i)\right] = H(Y_i|T_i) + \mathbb{E}_{T_i,W}\left[\mathrm{D_{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right] - I(W;Y_i|T_i). \tag{2}$$

Application: Controlling Label Information for KL Guided Marginal Alignment

Motivation: Discrepancy between $P_{Y|T}$ and $P_{Y'|T'}$

- Nguyen et al. (2022) shows that $\mathrm{D_{KL}}(P_{Y'|T'}||P_{Y|T}) \leq \mathrm{D_{KL}}(P_{Y'|X'}||P_{Y|X})$ if $I(X;Y) = I(T;Y)$. Penalizing the KL of the marginals is safe.

- The condition $I(X;Y) = I(T;Y)$ can be difficult to satisfy when ℓ is cross-entropy.
  By DPI on $Y - X - T$, $I(X;Y) \geq I(T;Y) = H(Y) - H(Y|T)$.

$$\mathbb{E}_{W,Z_i}\left[\ell(f_W(T_i),Y_i)\right] = H(Y_i|T_i) + \mathbb{E}_{T_i,W}\left[\mathrm{D_{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right] - I(W;Y_i|T_i). \quad (2)$$

Minimizing cross-entropy $\not\Rightarrow$ Minimizing $H(Y|T)$
$I(W;Y_i|T_i)$ increases $\Longrightarrow$ W just simply memorizes the label Y_i, resulting a form of overfitting.

Controlling Label Information

In Theorem 5, $I^{T'_j}(W;Z_i) = I^{T'_j}(W;T_i) + I^{T'_j}(W;Y_i|T_i)$.

Controlling Label Information

In Theorem 5, $I^{T_j'}(W;Z_i) = I^{T_j'}(W;T_i) + I^{T_j'}(W;Y_i|T_i)$.

Lemma 4 (Golden Formula)

For two random variables X and Y, we have

$$I(X;Y) = \inf_P \mathbb{E}_X\left[\mathrm{D_{KL}}(Q_{Y|X}||P)\right],$$

where the infimum is achieved at $P = Q_Y$.

Thus,

$$I^{T_j'}(W;Y_i|T_i) \leq \inf_Q \mathbb{E}_{T_i}\left[\mathrm{D_{KL}}(P_{W|Y_i,T_i,T_j'=t_j'}||Q_{W|T_i,T_j'=t_j'})\right].$$

Controlling Label Information

Assume $P = \mathcal{N}(W, \sigma^2 \mathrm{I}_d | Y_i, T_i, T'_j = t'_j)$ and let $Q = \mathcal{N}(\widetilde{W}, \tilde{\sigma}^2 \mathrm{I}_d | T_i, T'_j = t'_j)$, we have

$$I^{T'_j}(W; Y_i | T_i) \leq \mathbb{E}_{T_i}\left[\mathrm{D_{KL}}(P_{W|Y_i,T_i,T'_j=t'_j} || Q_{\tilde{W}|T_i,T'_j=t'_j})\right] \propto ||W - \widetilde{W}||^2.$$

Creating an auxiliary classifier $f_{\widetilde{w}}$ that does not depend on Y.

- In each iteration, we use the pseudo labels of target data (and source data) assigned by f_w to train $f_{\widetilde{w}}$
- Adding $||W - \widetilde{W}||^2$ as a regularizer in the training of W.

Experimental Results-RotatedMNIST

RotatedMNIST is built based on the MNIST dataset and consists of six domains, which are rotated MNIST images with rotation angle $0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ$ and 75°, respectively.

Table 1: RotatedMNIST.

Method	RotatedMNIST ($\mathbf{0}^\circ$ as source domain)					
	$\mathbf{15}^\circ$	$\mathbf{30}^\circ$	$\mathbf{45}^\circ$	$\mathbf{60}^\circ$	$\mathbf{75}^\circ$	**Ave**
ERM	97.5±0.2	84.1±0.8	53.9±0.7	34.2±0.4	22.3±0.5	58.4
DANN	97.3±0.4	90.6±1.1	68.7±4.2	30.8±0.6	19.0±0.6	61.3
MMD	97.5±0.1	95.3±0.4	73.6±2.1	44.2±1.8	32.1±2.1	68.6
CORAL	97.1±0.3	82.3±0.3	56.0±2.4	30.8±0.2	27.1±1.7	58.7
WD	96.7±0.3	93.1±1.2	64.1±3.3	41.4±7.6	27.6±2.0	64.6
KL	97.8±0.1	97.1±0.2	93.4±0.8	75.5±2.4	68.1±1.8	86.4
ERM-GP	97.5±0.1	86.2±0.5	62.0±1.9	34.8±2.1	26.1±1.2	61.2
ERM-CL	97.3±0.1	84.1±0.1	56.9±2.5	34.2±1.9	25.5±1.6	59.6
KL-GP	98.2±0.2	96.9±0.1	95.0±0.6	**88.0±8.1**	**78.1±2.5**	**91.2**
KL-CL	**98.4±0.2**	**97.3±0.2**	**95.6±0.1**	83.0±8.2	73.6±4.0	89.6

Experimental Results-Digits

Digits consists of 3 sub-datasets, namely MNIST, USPS and SVHN, and the corresponding domain adaptation tasks are **M$\to$U**, **U$\to$M**, **S$\to$M**.

Table 2: Digits.

	Digits			
Method	**M $\to$ U**	**U $\to$ M**	**S $\to$ M**	**Ave**
ERM	73.1±4.2	54.8±6.2	65.9±1.4	64.6
DANN	90.7±0.4	91.2±0.8	71.1±0.5	84.3
MMD	91.8±0.3	94.4±0.5	82.8±0.3	89.7
CORAL	88.0±1.9	83.3±0.1	69.3±0.6	80.2
WD	88.2±0.6	60.2±1.8	68.4±2.5	72.3
KL	98.2±0.2	97.3±0.5	92.5±0.9	96.0
ERM-GP	91.3±1.6	72.7±4.2	68.4±0.2	77.5
ERM-CL	88.9±0.4	71.2±3.6	73.5±1.4	77.9
KL-GP	98.8±0.1	**97.8±0.1**	**93.8±1.1**	**96.8**
KL-CL	**98.9±0.1**	97.7±0.1	93.0±0.3	96.5

Thank you!

zwang286@uottawa.ca