Implicit Class-Conditioned Domain Alignment for Unsupervised Domain Adaptation

Xiang Jiang[1,2] Qicheng Lao[1,4]
Stan Matwin[1,3] Mohammad Havaei[1]

[1]Imagia [2]Dalhousie University [3]Polish Academy of Sciences

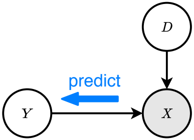
[4]Mila, Université de Montréal

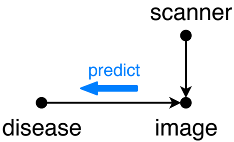
June 13, 2020

Introduction: Unsupervised Domain Adaptation (UDA)

The setup of UDA:

- observed variable X
- labeling function f, labels $Y = f(X)$
- domain variable D
- The goal is to learn $p(y|x)$ where
  - $\mathcal{D}_S = \{(x_i, f_S(x_i))\}_{i=1}^{n}$
  - $\mathcal{D}_T = \{x_j\}_{j=1}^{m}$
  - $f_S = f_T$





Related Work

Adversarial domain-discriminator based approaches [Ganin et al., 2016]:

$$\min_{\theta} \; \mathcal{L}(\mathcal{D}_S) + \lambda \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{1}$$

$$\max_{f} \; \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{2}$$

Related Work

Adversarial domain-discriminator based approaches [Ganin et al., 2016]:

$$\min_{\theta} \; \mathcal{L}(\mathcal{D}_S) + \lambda \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{1}$$

$$\max_{f} \; \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{2}$$

Limitation: $p_S(x) = p_T(x) \nRightarrow p_S(x|y) = p_T(x|y)$

Related Work

Adversarial domain-discriminator based approaches [Ganin et al., 2016]:

$$\min_{\theta} \ \mathcal{L}(\mathcal{D}_S) + \lambda \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{1}$$

$$\max_{f} \ \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{2}$$

Limitation: $p_S(x) = p_T(x) \not\Rightarrow p_S(x|y) = p_T(x|y)$

Prototype-based class-conditioned explicit alignment [Luo et al., 2017, Xie et al., 2018]:

$$\min_{\theta} \ \mathcal{L}(\mathcal{D}_S) + \lambda_1 \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) + \lambda_2 \mathcal{L}_{\text{explicit}} \tag{3}$$

$$\max_{f} \ \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{4}$$

where

$$\mathcal{L}_{\text{explicit}} = \mathbb{E}[\mathbf{c}_j{}^S - \mathbf{c}_j{}^T] \tag{5}$$

$$\mathbf{c}_j{}^S = \frac{1}{N_j} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_S} \mathbb{1}_{\{y_i = j\}} f_\phi(\mathbf{x}_i) \tag{6}$$

Related Work

Adversarial domain-discriminator based approaches [Ganin et al., 2016]:

$$\min_{\theta} \ \mathcal{L}(\mathcal{D}_S) + \lambda \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{1}$$

$$\max_{f} \ \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{2}$$

Limitation: $p_S(x) = p_T(x) \nRightarrow p_S(x|y) = p_T(x|y)$

Prototype-based class-conditioned explicit alignment [Luo et al., 2017, Xie et al., 2018]:

$$\min_{\theta} \ \mathcal{L}(\mathcal{D}_S) + \lambda_1 \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) + \lambda_2 \mathcal{L}_{\text{explicit}} \tag{3}$$

$$\max_{f} \ \mathrm{dis}(\mathcal{D}_S, \mathcal{D}_T) \tag{4}$$

where

$$\mathcal{L}_{\text{explicit}} = \mathbb{E}[\mathbf{c}_j{}^S - \mathbf{c}_j{}^T] \tag{5}$$

$$\mathbf{c}_j{}^S = \frac{1}{N_j} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_S} \mathbb{1}_{\{y_i = j\}} f_\phi(\mathbf{x}_i) \tag{6}$$

Limitation: Error accumulation in explicit optimization on pseudo-labels

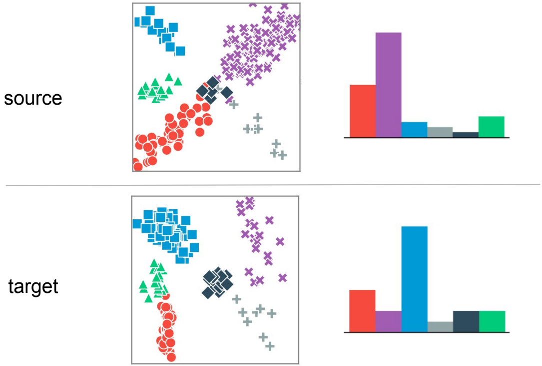
Motivations

- Applied motivation
- Theoretical motivation

Applied Motivation

Challenges for applying UDA in real-world applications [Tan et al., 2019]:

- *within-domain* class imbalance;
- *between-domain* class distribution shift, aka, prior probability shift.



Theoretical Motivation: Empirical Domain Divergence

Definition ([Ben-David et al., 2010])

The $\mathcal{H}\Delta\mathcal{H}$ divergence between two domains is defined as

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h,h' \in \mathcal{H}} \left| \mathbb{E}_{\mathcal{D}_T}[h \neq h'] - \mathbb{E}_{\mathcal{D}_S}[h \neq h'] \right|, \tag{7}$$

Definition (mini-batch based empirical domain discrepancy)

Let $\mathcal{B}_S$, $\mathcal{B}_T$ be minibatches from $\mathcal{U}_S$ and $\mathcal{U}_T$, respectively, where $\mathcal{B}_S \subseteq \mathcal{U}_S$, $\mathcal{B}_T \subseteq \mathcal{U}_T$, and $|\mathcal{B}_S| = |\mathcal{B}_T|$. The empirical estimation of $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{B}_S, \mathcal{B}_T)$ over the minibatches $\mathcal{B}_S$, $\mathcal{B}_T$ is defined as

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{B}_S, \mathcal{B}_T) = \sup_{h,h' \in \mathcal{H}} \left| \sum_{\mathcal{B}_T} [h \neq h'] - \sum_{\mathcal{B}_S} [h \neq h'] \right|. \tag{8}$$

Theoretical Motivation: The Decomposition

Theorem (The decomposition of $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{B}_S, \mathcal{B}_T)$)

We define three disjoint sets on the label space: $Y_C := Y_S \cap Y_T$, $\overline{Y_S} := Y_S - Y_C$, *and* $\overline{Y_T} := Y_T - Y_C$. *We also define the following disjoint sets on the input space where* $\mathcal{B}_S^C := \{x \in \mathcal{B}_S \mid y \in Y_C\}$, $\mathcal{B}_S^{\overline{C}} := \{x \in \mathcal{B}_S \mid y \notin Y_C\}$, $\mathcal{B}_T^C := \{x \in \mathcal{B}_T \mid y \in Y_C\}$, $\mathcal{B}_T^{\overline{C}} := \{x \in \mathcal{B}_T \mid y \notin Y_C\}$. *The empirical* $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{B}_S, \mathcal{B}_T)$ *divergence can be decomposed into as the following:*

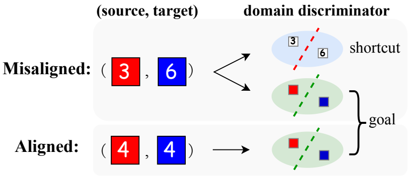
$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{B}_S, \mathcal{B}_T) = \sup_{h, h' \in \mathcal{H}} \left| \xi^C(h, h') + \xi^{\overline{C}}(h, h') \right|, \tag{9}$$

where

$$\xi^C(h, h') = \sum_{\mathcal{B}_T^C} \mathbb{1}\left[h \neq h'\right] - \sum_{\mathcal{B}_S^C} \mathbb{1}\left[h \neq h'\right], \tag{10}$$

$$\xi^{\overline{C}}(h, h') = \sum_{\mathcal{B}_T^{\overline{C}}} \mathbb{1}\left[h \neq h'\right] - \sum_{\mathcal{B}_S^{\overline{C}}} \mathbb{1}\left[h \neq h'\right]. \tag{11}$$

Theoretical Motivation: Domain-Discriminator Shortcut



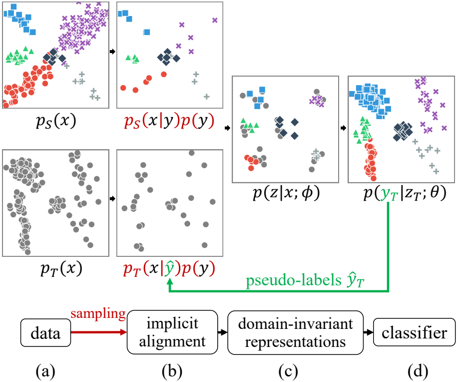
Remark (The domain discriminator shortcut)

Let f_c be a classifier that maps x to a class label y_c. Let f_d be a domain discriminator that maps x to a binary domain label y_d. For the empirical class-misaligned divergence $\xi^{\overline{C}}(h, h')$ with sample $x \in \mathcal{B}_S^{\overline{C}} \cup \mathcal{B}_T^{\overline{C}}$, there exists a domain discriminator shortcut function

$$f_d(x) = \begin{cases} 1 & f_c(x) \in \overline{Y_S} \\ 0 & f_c(x) \in \overline{Y_T}, \end{cases} \tag{12}$$

such that the domain label can be solely determined by the domain-specific class labels. (More pronounced under imbalance and distribution shift.)

Proposed Approach



- For $p_S(x)$, we sample $x \sim p_S(x|y)p(y)$ based on the *alignment distribution* $p(y)$
- For $p_T(x)$, we sample a *class aligned* minibatch $x \sim p_T(x|\hat{y})p(y)$ using identical $p(y)$, with the help of pseudo-labels $\hat{y}_T$

Proposed Approach

Input: dataset $S = \{(x_i, y_i)\}_{i=1}^{N}$, $T = \{x_i\}_{i=1}^{M}$,
label space $\mathcal{Y}$, label alignment distribution $p(y)$,
classifier $f_c(\cdot; \theta)$
while not converged **do**
  # predict pseudo-labels for T
  $\hat{T} \leftarrow \{(x_i, \hat{y}_i)\}_{i=1}^{M}$ where $x_i \in T$ and $\hat{y}_i = f_c(x_i; \theta)$
  # sample N unique classes in the label space
  $Y \leftarrow$ draw N samples in $\mathcal{Y}$ from $p(y)$
  # sample K examples conditioned on each $y_j \in Y$
  for y_j in Y **do**
    $(X'_S, Y'_S) \leftarrow$draw K samples in S from $p_S(x|y = y_j)$
    $X'_T \leftarrow$draw K samples in $\hat{T}$ from $p_T(x|\hat{y} = y_j)$
  end for
  # domain adaptation training on this minibatch
  train minibatch (X'_S, Y'_S, X'_T)
end while

Advantages of the proposed approach

1. Minimizes the class-misaligned divergence $\xi^{\overline{C}}(h, h')$, providing a more reliable empirical estimation of domain divergence;

Advantages of the proposed approach

1. Minimizes the class-misaligned divergence $\xi^{\overline{C}}(h, h')$, providing a more reliable empirical estimation of domain divergence;
2. Provides balanced training across all classes;

Advantages of the proposed approach

1. Minimizes the class-misaligned divergence $\xi^{\overline{C}}(h, h')$, providing a more reliable empirical estimation of domain divergence;
2. Provides balanced training across all classes;
3. Removes the need to optimize model parameters from pseudo-labels explicitly;

Advantages of the proposed approach

1. Minimizes the class-misaligned divergence $\xi^{\overline{C}}(h, h')$, providing a more reliable empirical estimation of domain divergence;
2. Provides balanced training across all classes;
3. Removes the need to optimize model parameters from pseudo-labels explicitly;
4. Simple to implement and is orthogonal to different domain discrepancy measures: DANN and MDD.

Extending Implicit Alignment to MDD

MDD is defined as

$$d_{f,\mathcal{F}}(S,T) = \sup_{f' \in \mathcal{F}} \Big(\mathrm{disp}_{\mathcal{D}_T}(f', f) - \mathrm{disp}_{\mathcal{D}_S}(f', f) \Big), \tag{13}$$

where f and f' are two independent scoring functions that predict class probabilities, and $\mathrm{disp}(f', f)$ is a disparity measure between the scores provided by the classifiers f' and f.

We introduce a masking scheme on f and f' defined as

$$\begin{aligned} &\hat{d}_{f,\mathcal{F}}(\mathcal{B}_S, \mathcal{B}_T) \\ &= \sup_{f' \in \mathcal{F}} \Big(\sum_{\mathcal{B}_T} \mathrm{disp}(f' \odot \omega, f \odot \omega) - \sum_{\mathcal{B}_S} \mathrm{disp}(f' \odot \omega, f \odot \omega) \Big), \end{aligned} \tag{14}$$

where $f \odot \omega$ denotes element-wise multiplication between the output of f and ω. The alignment mask ω is a binary vector that denotes whether the i-th class is present in the sampled classes Y (i.e., the classes that we intend to align in the current minibatch).

Experiment Setup

Datasets:

- Office-31 [Saenko et al., 2010]
- Office-Home [Venkateswara et al., 2017]
  1. standard [Venkateswara et al., 2017]: natrual imbalance
  2. balanced [Tan et al., 2019]
  3. "RS-UT" [Tan et al., 2019]
- VisDA2017 (synthetic→real) [Peng et al., 2017]
- MNIST and SVHN (ablation studies)

Baselines:

- Covariate and Label Shift CO-ALignment (COAL) [Tan et al., 2019]
- Explicit alignment [Liang et al., 2019b, Liang et al., 2019a]

PyTorch Code: `https://github.com/xiangdal/implicit_alignment`

Dataset Statistics

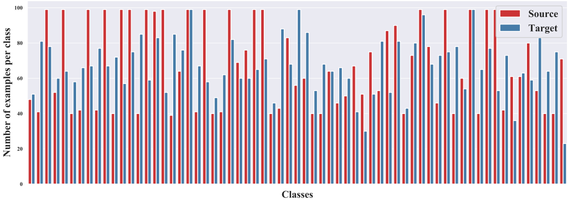


Figure: Class frequency of Cl→Rw, Office-Home (standard)

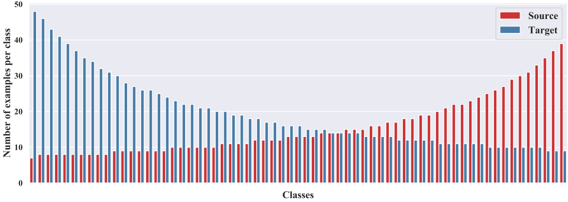


Figure: Class frequency of of Cl→Rw, Office-Home (RS-UT)

Empirical Results: Office-Home (RS-UT)

Methods	Rw→Pr	Rw→Cl	Pr→Rw	Pr→Cl	Cl→Rw	Cl→Pr	Avg
Source Only†	69.77	38.35	67.31	35.84	53.31	52.27	52.81
BSP [Chen et al., 2019]†	72.80	23.82	66.19	20.05	32.59	30.36	40.97
PADA [Cao et al., 2018]†	60.77	32.28	57.09	26.76	40.71	38.34	42.66
BBSE [Lipton et al., 2018]†	61.10	33.27	62.66	31.15	39.70	38.08	44.33
MCD [Saito et al., 2018]†	66.03	33.17	62.95	29.99	44.47	39.01	45.94
DAN [Long et al., 2015]†	69.35	40.84	66.93	34.66	53.55	52.09	52.90
F-DANN [Wu et al., 2019]†	68.56	40.57	67.32	37.33	55.84	53.67	53.88
JAN [Long et al., 2017]†	67.20	43.60	68.87	39.21	57.98	48.57	54.24
DANN [Ganin et al., 2016]†	71.62	46.51	68.40	38.07	58.83	58.05	56.91
MDD (random sampler)	71.21	44.78	69.31	42.56	52.10	52.70	55.44
MDD (source-balanced sampler)	76.06	47.38	71.56	40.03	57.46	58.54	58.50
COAL [Tan et al., 2019]†,‡	73.65	42.58	73.26	40.61	59.22	57.33	58.40
MDD+Explicit Alignment (basic)‡	69.52	44.70	69.59	40.27	53.02	53.39	55.08
MDD+Explicit Alignment (moving avg.)‡	71.37	45.26	69.69	40.28	52.92	52.69	55.37
MDD+Explicit Alignment (curriculum)‡	70.02	45.48	69.71	40.86	53.26	52.99	55.39
MDD+Implicit Alignment	**76.08**	**50.04**	**74.21**	**45.38**	**61.15**	**63.15**	**61.67**

† *Source*: Data of these baseline methods are cited from [Tan et al., 2019].
‡ Methods using explicit class-conditioned domain alignment.

Empirical Results: Office-31 (standard)

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
Source only	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DAN [Long et al., 2015]	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
DANN [Ganin et al., 2016]	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
ADDA [Tzeng et al., 2017]	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
JAN [Long et al., 2017]	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
MADA [Pei et al., 2018]	90.0 ± 0.1	97.4±0.1	99.6±0.1	87.8±0.2	70.3±0.3	66.4±0.3	85.2
GTA [Sankaranarayanan et al., 2018]	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
MCD [Saito et al., 2018]	88.6±0.2	98.5±0.1	**100.0**±.0	92.2±0.2	69.5±0.1	69.7±0.3	86.5
CDAN [Long et al., 2018]	94.1±0.1	98.6±0.1	**100.0**±.0	92.9±0.2	71.0±0.3	69.3±0.3	87.7
MDD [Zhang et al., 2019]	**94.5**±0.3	98.4±0.1	**100.0**±.0	**93.5**±0.2	74.6±0.3	72.2±0.1	**88.9**
PACET [Liang et al., 2019b]‡	90.8	97.6	99.8	90.8	73.5	73.6	87.4
CAT [Deng et al., 2019]‡	94.4±0.1	98.0±0.2	**100.0**±0.0	90.8±1.8	72.2±0.2	70.2±0.1	87.6
MDD (source-balanced sampler)	90.4±0.4	**98.7**±0.1	99.9±0.1	90.4±0.2	75.0±0.5	73.7±0.9	88.0
MDD+Explicit Alignment‡	92.3±0.1	98.2±0.1	99.8±.0	92.3±0.3	74.6±0.2	72.9±0.7	88.4
MDD+Implicit Alignment	90.3±0.2	**98.7**±0.1	99.8±.0	92.1±0.5	**75.3**±0.2	**74.9**±0.3	88.8

‡ Methods using explicit class-conditioned domain alignment.

Empirical Results: Office-Home (standard)

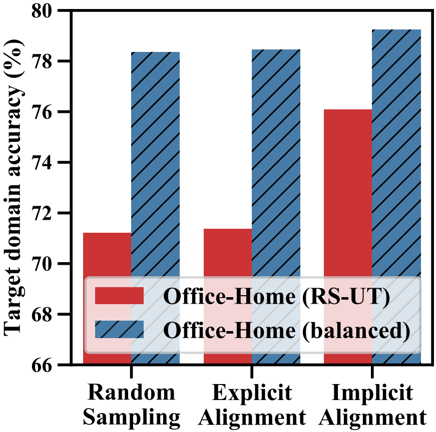
Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
Source only	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [Long et al., 2015]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [Ganin et al., 2016]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [Long et al., 2017]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN [Long et al., 2018]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
BSP [Chen et al., 2019]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
MDD [Zhang et al., 2019]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	**72.5**	**60.2**	82.3	68.1
MCS [Liang et al., 2019a]‡	55.9	73.8	79.0	57.5	69.9	71.3	58.4	50.3	78.2	65.9	53.2	82.2	66.3
MDD+Explicit Alignment‡	54.3	74.6	77.6	60.7	71.9	71.4	62.1	52.4	76.9	71.1	57.6	81.3	67.7
MDD (source-balanced sampler)	55.3	75.0	79.1	62.3	70.1	73.2	63.5	53.2	78.7	70.4	56.2	82.0	68.3
MDD+Implicit Alignment	**56.2**	**77.9**	**79.2**	**64.4**	**73.1**	**74.4**	**64.2**	**54.2**	**79.9**	71.2	58.1	**83.1**	**69.5**

‡ Methods using explicit class-conditioned domain alignment.

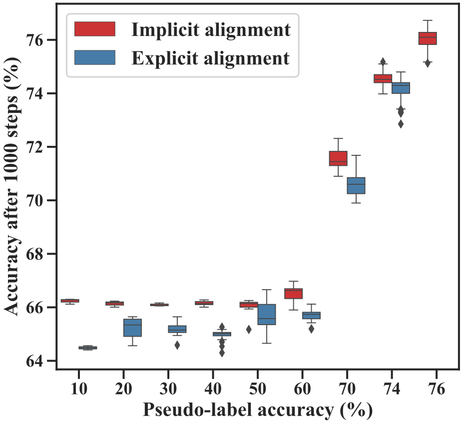
Empirical Results: VisDA2017

method	acc. (%)
JAN [Long et al., 2017]	61.6
GTA[Sankaranarayanan et al., 2018]	69.5
MCD [Saito et al., 2018]	69.8
CDAN [Long et al., 2018]	70.0
MDD [Zhang et al., 2019]	74.6
MDD+Explicit Alignment	67.1
MDD+Implicit Alignment	**75.8**

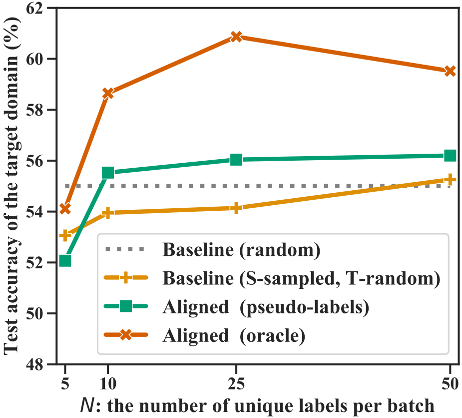
Ablation Studies: Implicit vs. Explicit Alignment



Ablation Studies: Robustness to Pseudo-label Errors



Ablation Studies: Class Diversity and Alignment



Interactions between class imbalance and distribution shift

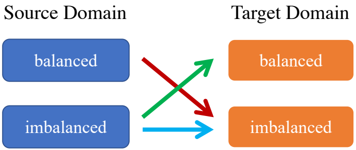


Table: S-balanced, T-imbalanced.

method	SVHN→MNIST mild	SVHN→MNIST extreme	MNIST→SVHN mild	MNIST→SVHN extreme
source only	67.4±7.3	66.3±3.3	**32.5**±2.9	**28.2**±2.3
DANN	78.2±2.8	59.1±0.8	20.9±6.0	20.5±3.1
DANN+implicit	**88.6**±0.7	**82.2**±2.1	**32.4**±2.1	**28.9**±3.3

Table: S-imbalanced, T-balanced.

method	SVHN→MNIST mild	SVHN→MNIST extreme	MNIST→SVHN mild	MNIST→SVHN extreme
source only	65.2±2.1	53.3±1.3	**31.6**±3.3	**32.8**±0.9
DANN	82.0±0.7	52.3±2.3	23.4±3.6	25.9±0.5
DANN+implicit	**91.0**±1.9	**87.1**±2.6	**34.9**±0.5	**31.1**±2.9

Table: Both domains imbalanced.

method	SVHN→MNIST mild	SVHN→MNIST extreme	MNIST→SVHN mild	MNIST→SVHN extreme
source only	60.9±5.2	51.2±5.9	30.6±1.3	**27.1**±1.7
DANN	67.6±0.8	40.5±5.5	23.4±1.6	18.8±2.9
DANN+implicit	**88.6**±0.6	**70.5**±3.6	**36.3**±2.5	**27.9**±2.4

Conclusion

- We introduce an implicit class-conditioned domain alignment approach;

Conclusion

- We introduce an implicit class-conditioned domain alignment approach;
- A more reliable measure of empirical domain divergence;

Conclusion

- We introduce an implicit class-conditioned domain alignment approach;
- A more reliable measure of empirical domain divergence;
- Implicit alignment works well under extreme within-domain class imbalance and between-domain class distribution shift, as well as competitive results on standard UDA tasks;

Conclusion

- We introduce an implicit class-conditioned domain alignment approach;
- A more reliable measure of empirical domain divergence;
- Implicit alignment works well under extreme within-domain class imbalance and between-domain class distribution shift, as well as competitive results on standard UDA tasks;
- The proposed approach is simple to implement and orthogonal to the choice of domain adaptation algorithms.

Future Work

- Other domain adaptation setups, e.g., open set domain adaptation and partial domain adaptation.

Future Work

- Other domain adaptation setups, e.g., open set domain adaptation and partial domain adaptation.
- Cost-sensitive learning for domain adaptation.

Future Work

- Other domain adaptation setups, e.g., open set domain adaptation and partial domain adaptation.
- Cost-sensitive learning for domain adaptation.
- More work on domain adaptation in the presence of within-domain imbalance and between-domain class distribution shift are needed to facilitate safer use of machine learning models in the real-world.

Thank you!

References I

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010).
A theory of learning from different domains.
Machine learning, 79(1-2):151–175.

Cao, Z., Ma, L., Long, M., and Wang, J. (2018).
Partial adversarial domain adaptation.
In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150.

Chen, X., Wang, S., Long, M., and Wang, J. (2019).
Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation.
In *International Conference on Machine Learning*, pages 1081–1090.

References II

Deng, Z., Luo, Y., and Zhu, J. (2019).
Cluster alignment with a teacher for unsupervised domain adaptation.
In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016).
Domain-adversarial training of neural networks.
The Journal of Machine Learning Research, 17(1):2096–2030.

Liang, J., He, R., Sun, Z., and Tan, T. (2019a).
Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2975–2984.

References III

Liang, J., He, R., Sun, Z., and Tan, T. (2019b).
Exploring uncertainty in pseudo-label guided unsupervised domain adaptation.
Pattern Recognition, 96:106996.

Lipton, Z. C., Wang, Y.-X., and Smola, A. (2018).
Detecting and correcting for label shift with black box predictors.
arXiv preprint arXiv:1802.03916.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015).
Learning transferable features with deep adaptation networks.
In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 97–105. JMLR. org.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2018).
Conditional adversarial domain adaptation.
In *Advances in Neural Information Processing Systems*, pages 1640–1650.

References IV

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017).
Deep transfer learning with joint adaptation networks.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org.

Luo, Z., Zou, Y., Hoffman, J., and Fei-Fei, L. F. (2017).
Label efficient learning of transferable representations acrosss domains and tasks.
In *Advances in Neural Information Processing Systems*, pages 165–177.

Pei, Z., Cao, Z., Long, M., and Wang, J. (2018).
Multi-adversarial domain adaptation.
In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. (2017).
Visda: The visual domain adaptation challenge.
arXiv preprint arXiv:1710.06924.

References V

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. (2010).
Adapting visual category models to new domains.
In *European conference on computer vision*, pages 213–226. Springer.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018).
Maximum classifier discrepancy for unsupervised domain adaptation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. (2018).
Generate to adapt: Aligning domains using generative adversarial networks.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512.

Tan, S., Peng, X., and Saenko, K. (2019).
Generalized domain adaptation with covariate and label shift co-alignment.
arXiv preprint arXiv:1910.10320.

References VI

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017).
Adversarial discriminative domain adaptation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017).
Deep hashing network for unsupervised domain adaptation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. (2019).
Domain adaptation with asymmetrically-relaxed distribution alignment.
arXiv preprint arXiv:1903.01689.

Xie, S., Zheng, Z., Chen, L., and Chen, C. (2018).
Learning semantic representations for unsupervised domain adaptation.
In *International Conference on Machine Learning*, pages 5419–5428.

References VII

Zhang, Y., Liu, T., Long, M., and Jordan, M. (2019).
Bridging theory and algorithm for domain adaptation.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7404–7413, Long Beach, California, USA. PMLR.