

Domain Adaptation: Theory, Algorithms, and Open Library

Mingsheng Long

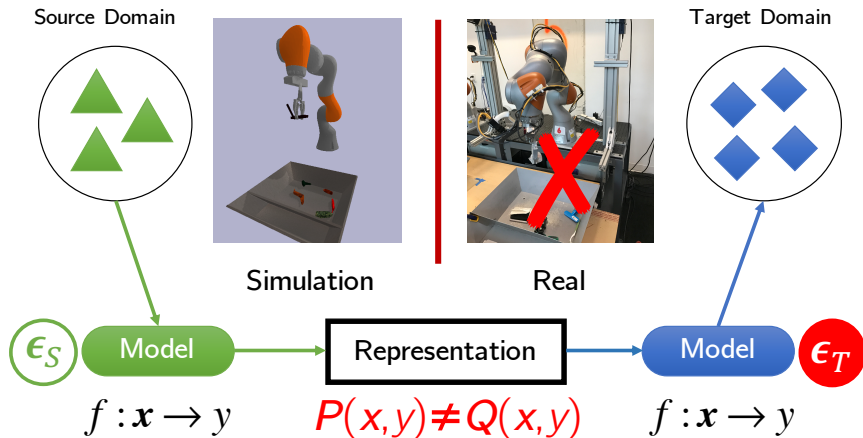
School of Software, Tsinghua University
National Engineering Laboratory for Big Data Software

mingsheng@tsinghua.edu.cn

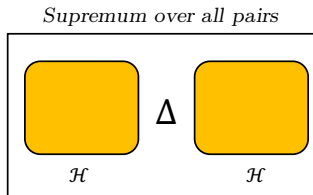
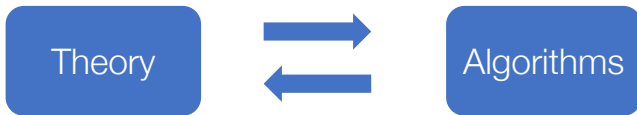
Workshop on Distribution Shifts, NeurIPS 2022

Domain Adaptation

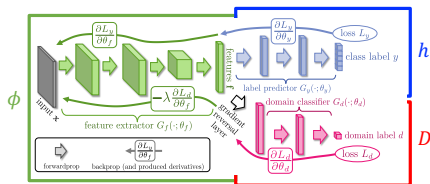
- Machine learning across domains of different distributions $P \neq Q$
- OOD: Out-of-Distribution** (from IID to OOD)
- How to bound **generalization error** on target domain for OOD case?



How to Bridge Theory and Algorithms?



$\mathcal{H}\Delta\mathcal{H}$ -Divergence



There is nothing more practical than a good theory.

—Vladimir Vapnik

Outline

1 Domain Adaptation

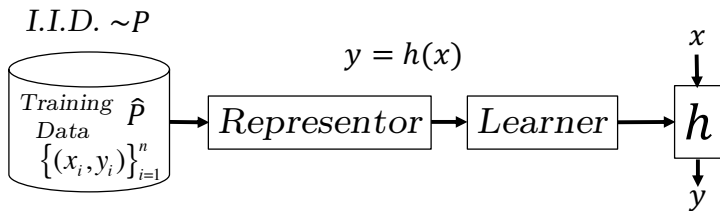
2 Theory and Algorithms

- Classic Theory
- Margin Theory
- Localization Theory

3 Open Library

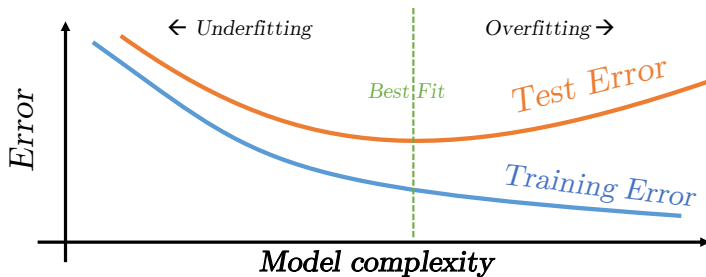
- TLLib: Transfer Learning Library

Statistical Learning



- Classification problem with **01-loss** $[\cdot \neq \cdot]$.
- **Training error**: $\epsilon_{\hat{P}}(h) = \frac{1}{n} \sum_{i=1}^n [h(\mathbf{x}_i) \neq y_i] = \mathbb{E}_{(\mathbf{x}, y) \sim \hat{P}} [h(\mathbf{x}) \neq y]$.
- **Test error**: $\epsilon_P(h) = \mathbb{E}_{(\mathbf{x}, y) \sim P} [h(\mathbf{x}) \neq y]$.
- Can we control $\epsilon_P(h)$ with observable $\epsilon_{\hat{P}}(h)$?

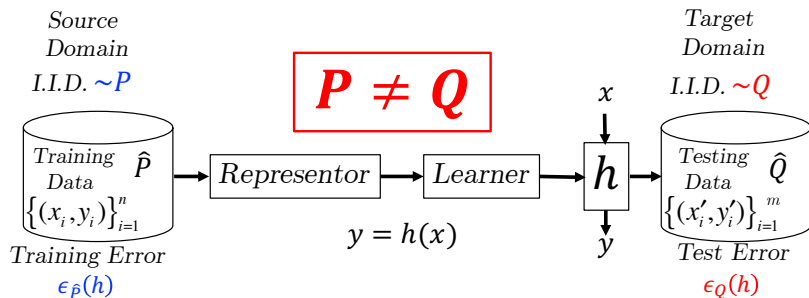
Statistical Learning Theory



- Generalization error depends on sample size n and **model complexity**.
- For hypothesis space \mathcal{H} with **VC-dimension** d , we have bound:

$$\epsilon_P(h) \leq \epsilon_{\hat{P}}(h) + O\left(\sqrt{\frac{d \log n + \log \frac{2}{\delta}}{n}}\right)$$

Domain Adaptation



- Labeled data of size n sampled from a source domain P .
- Unlabeled data of size m sampled from a **different** target domain Q .
- Can we control target error $\epsilon_Q(h)$ with observable $\epsilon_{\hat{P}}(h)$?
 - **Disparity on D** : $\epsilon_D(h_1, h_2) = \mathbb{E}_{(x,y) \sim D} [h_1(x) \neq h_2(x)]$.
 - Why use it? Computation of disparity **does not require** (target) label!

Relating Target Risk to Source Risk

Theorem (Bound with Disparity)

For domain adaptation classification tasks, define the *ideal joint hypothesis* as $h^* = \arg \min_{h \in \mathcal{H}} [\epsilon_P(h) + \epsilon_Q(h)]$, the target risk $\epsilon_Q(h)$ can be bounded by the source risk $\epsilon_P(h)$, the *ideal joint error*, and the *disparity difference*:

$$\epsilon_Q(h) \leq \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)| \quad (1)$$

Proof.

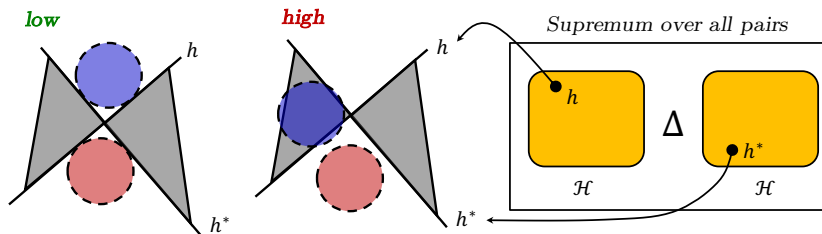
Simply using the *triangle inequalities* of the 01-loss, we have

$$\begin{aligned} \epsilon_Q(h) &\leq \epsilon_Q(h^*) + \epsilon_Q(h, h^*) \\ &= \epsilon_Q(h^*) + \epsilon_P(h, h^*) + \epsilon_Q(h, h^*) - \epsilon_P(h, h^*) \\ &\leq \epsilon_Q(h^*) + \epsilon_P(h, h^*) + |\epsilon_Q(h, h^*) - \epsilon_P(h, h^*)| \\ &\leq \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)| \end{aligned} \quad (2)$$



$\mathcal{H}\Delta\mathcal{H}$ -Divergence¹

- **Assumption:** Small ideal joint error $\epsilon_{ideal} = \epsilon_P(h^*) + \epsilon_Q(h^*)$.
- We can illustrate the **disparity difference** $|\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)|$:



- However, h^* is unknown and h is undefined. Consider **worse-case!**
- **$\mathcal{H}\Delta\mathcal{H}$ -Divergence:** $d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) \triangleq \sup_{h, h' \in \mathcal{H}} |\epsilon_P(h, h') - \epsilon_Q(h, h')|$
- Can be estimated from **finite unlabeled** samples of source and target.

¹Ben-David et al. *A Theory of Learning from Different Domains*. Machine Learning, 2010.

Bound $\mathcal{H}\Delta\mathcal{H}$ -Divergence with Domain Discriminator

Theorem (Generalization Bound with $\mathcal{H}\Delta\mathcal{H}$ -Divergence)

Denote by d the VC-dimension of hypothesis space \mathcal{H} . We have

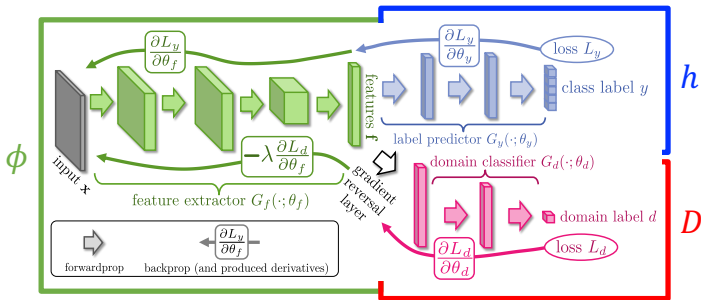
$$\epsilon_Q(h) \leq \epsilon_{\hat{P}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q}) + \epsilon_{ideal} + O\left(\sqrt{\frac{d \log n}{n}} + \sqrt{\frac{d \log m}{m}}\right) \quad (3)$$

- However, $\mathcal{H}\Delta\mathcal{H}$ -Divergence is **hard to compute and optimize**.
- For **binary** hypothesis h , $\mathcal{H}\Delta\mathcal{H}$ -Divergence can be further bounded by

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) &\triangleq \sup_{h, h' \in \mathcal{H}} |\epsilon_P(h, h') - \epsilon_Q(h, h')| \\ &= \sup_{\delta \in \mathcal{H}\Delta\mathcal{H}} |\mathbb{E}_P[\delta(\mathbf{x}) \neq 0] - \mathbb{E}_Q[\delta(\mathbf{x}) \neq 0]| \\ &\leq \sup_{D \in \mathcal{H}_D} |\mathbb{E}_P[D(\mathbf{x}) = 1] + \mathbb{E}_Q[D(\mathbf{x}) = 0]| \end{aligned} \quad (4)$$

- This bound can be estimated by training a **domain discriminator $D(\mathbf{x})$** .

Domain Adversarial Neural Network (DANN)²



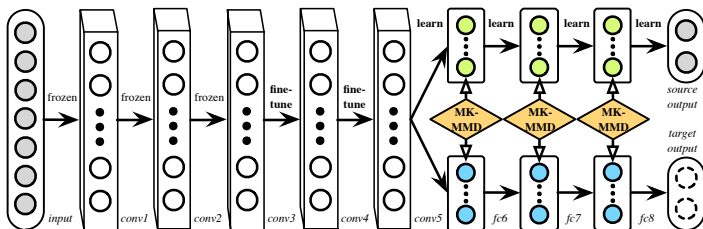
Adversarial domain adaptation: learn ϕ to minimize $d_{\mathcal{H}\Delta\mathcal{H}}(\phi(P), \phi(Q))$.

$$\min_{\phi, h} \left\{ \mathbb{E}_{(x, y) \sim P} L(h(\phi(x)), y) + \lambda \max_D (\mathbb{E}_P L(D(\phi(x)), 1) + \mathbb{E}_Q L(D(\phi(x)), 0)) \right\} \quad (5)$$

Supervised Learning on source + Upper-Bound of $d_{\mathcal{H}\Delta\mathcal{H}}$ on source/target

²Ganin et al. Domain Adversarial Training of Neural Networks. JMLR 2016.

Deep Adaptation Network (DAN)³



Optimal domain matching: yield upper-bound by multiple kernel learning

$$d_k^2(P, Q) \triangleq \|\mathbf{E}_P[\phi(\mathbf{x}^s)] - \mathbf{E}_Q[\phi(\mathbf{x}^t)]\|_{\mathcal{H}_k}^2 \quad (6)$$

$$\min_{\theta \in \Theta} \frac{1}{n_s} \sum_{i=1}^{n_s} L(\theta(\mathbf{x}_i^s), y_i^s) + \lambda \max_{k \in \mathcal{K}} \sum_{\ell=1}^{l_2} d_k^2(\hat{P}_\ell, \hat{Q}_\ell) \quad (7)$$

Works better than f -Divergences when domains are less overlapping

³Long et al. *Learning Transferable Features with Deep Adaptation Networks*. ICML 2015.

Outline

1 Domain Adaptation

2 Theory and Algorithms

- Classic Theory
- **Margin Theory**
- Localization Theory

3 Open Library

- TLLib: Transfer Learning Library

Theory vs. Practice



- **Theory** vs. **Practice**:
- Binary Classification vs. Multiclass Classification.
- Discrete Classifier vs. Classifier with Scoring Function.
- $d_{\mathcal{H}\Delta\mathcal{H}}$'s bound is minimized vs. $d_{\mathcal{H}\Delta\mathcal{H}}$ is hard to estimate & optimize.
- How to bridge the gap between theory and algorithm?

Step I: Disparity Discrepancy (DD)⁴

Definition (Disparity Discrepancy (DD))

Given a hypothesis space \mathcal{H} and a *specific hypothesis* $h \in \mathcal{H}$, the Disparity Discrepancy (DD) is

$$d_{h,\mathcal{H}}(P, Q) = \sup_{h' \in \mathcal{H}} (\mathbb{E}_Q[h' \neq h] - \mathbb{E}_P[h' \neq h]) \quad (8)$$

Theorem (Bound with Disparity Discrepancy)

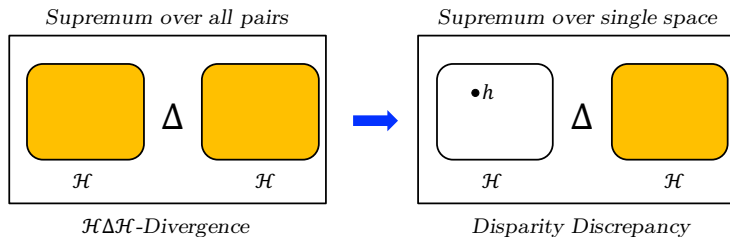
For any $\delta > 0$ and binary classifier $h \in \mathcal{H}$, with probability $1 - 3\delta$, we have

$$\begin{aligned} \epsilon_Q(h) &\leq \epsilon_{\hat{P}}(h) + d_{h,\mathcal{H}}(\hat{P}, \hat{Q}) + \epsilon_{ideal} + 2\mathfrak{R}_{n,P}(\mathcal{H}\Delta\mathcal{H}) \\ &\quad + 2\mathfrak{R}_{n,P}(\mathcal{H}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} + 2\mathfrak{R}_{m,Q}(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \end{aligned} \quad (9)$$

⁴Zhang & Long. *Bridging Theory and Algorithm for Domain Adaptation*. ICML 2019.

Step I: Disparity Discrepancy (DD)

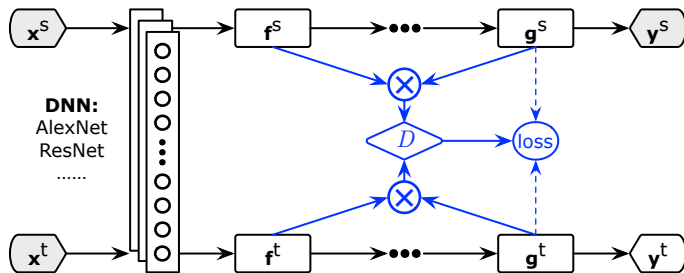
- Disparity Discrepancy (DD) is **tighter** than $\mathcal{H}\Delta\mathcal{H}$ -Divergence.



- DD can have connections to **conditional domain discriminator** $D(\mathbf{x}, h(\mathbf{x}))$.

$$\begin{aligned}
 d_{h,\mathcal{H}}(P, Q) &\triangleq \sup_{h' \in \mathcal{H}} (\epsilon_P(h, h') - \epsilon_Q(h, h')) \\
 &= \sup_{h' \in \mathcal{H}} (\mathbb{E}_P[|h(\mathbf{x}) - h'(\mathbf{x})| \neq 0] - \mathbb{E}_Q[|h(\mathbf{x}) - h'(\mathbf{x})| \neq 0]) \quad (10) \\
 &\leq \sup_{D \in \mathcal{H}_D} (\mathbb{E}_P[D(\mathbf{x}, h(\mathbf{x})) = 1] + \mathbb{E}_Q[D(\mathbf{x}, h(\mathbf{x})) = 0])
 \end{aligned}$$

Conditional Domain Adversarial Network (CDAN)⁵



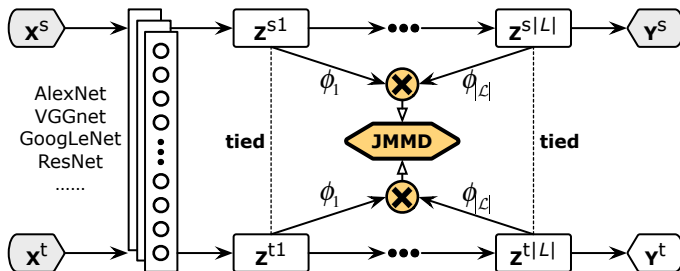
Conditional adversarial domain adaptation: minimize $d_{h,\mathcal{H}}(\phi(P), \phi(Q))$.

$$\begin{aligned} \min_G \mathcal{E}(G) - \lambda \mathcal{E}(D, G) \\ \min_D \mathcal{E}(D, G), \end{aligned} \quad (11)$$

$$\mathcal{E}(D, G) = -\mathbb{E}_{\mathbf{x}_i^s \sim \mathcal{D}_s} \log [D(\mathbf{f}_i^s \otimes \mathbf{g}_i^s)] - \mathbb{E}_{\mathbf{x}_j^t \sim \mathcal{D}_t} \log [1 - D(\mathbf{f}_j^t \otimes \mathbf{g}_j^t)] \quad (12)$$

⁵Long et al. Conditional Adversarial Domain Adaptation. NIPS 2018.

Joint Adaptation Network (JAN)⁶



Joint distribution matching: cross-covariance of multiple random vectors

$$d_k^2(P, Q) \triangleq \left\| \mathbf{E}_P \left[\bigotimes_{\ell=1}^m \phi_{\ell}(\mathbf{x}_{\ell}^s) \right] - \mathbf{E}_Q \left[\bigotimes_{\ell=1}^m \phi_{\ell}(\mathbf{x}_{\ell}^t) \right] \right\|_{\mathcal{H}_k}^2 \quad (13)$$

$$\min_{\theta \in \Theta} \max_{k \in \mathcal{K}} \frac{1}{n_a} \sum_{i=1}^{n_a} L(\theta(\mathbf{x}_i^a), y_i^a) + \lambda d_k^2(\hat{P}_{1:l_2}, \hat{Q}_{1:l_2}) \quad (14)$$

Works better than f -Divergences when domains are less overlapping

⁶Long et al. Deep Transfer Learning with Joint Adaptation Networks. ICML 2017.

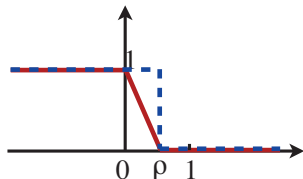
Multiclass Classification Formulation

- **Scoring function:** $f \in \mathcal{F} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$
- **Labeling function** induced by f : $h_f : \mathbf{x} \mapsto \arg \max_{y \in \mathcal{Y}} f(\mathbf{x}, y)$
- **Labeling function class:** $\mathcal{H} = \{h_f | f \in \mathcal{F}\}$
- **Margin** of a hypothesis f :

$$\rho_f(\mathbf{x}, y) = \frac{1}{2}(f(\mathbf{x}, y) - \max_{y' \neq y} f(\mathbf{x}, y'))$$

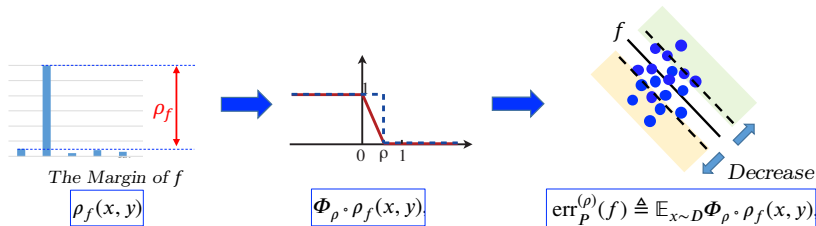
- **Margin Loss:**

$$\Phi_{\rho}(\mathbf{x}) = \begin{cases} 0 & \rho \leq \mathbf{x} \\ 1 - \mathbf{x}/\rho & 0 \leq \mathbf{x} \leq \rho \\ 1 & \mathbf{x} \leq 0 \end{cases}$$



Margin Theory

- Margin error: $\epsilon_D^{(\rho)}(f) = \mathbb{E}_{(\mathbf{x}, y) \sim D} [\Phi_\rho(\rho_f(\mathbf{x}, y))]$
- This error takes the **margin** of the hypothesis f into consideration:



- Given a class of scoring functions \mathcal{F} , $\Pi_1 \mathcal{F}$ is defined as

$$\Pi_1 \mathcal{F} = \{\mathbf{x} \mapsto f(\mathbf{x}, y) \mid y \in \mathcal{Y}, f \in \mathcal{F}\}. \quad (15)$$

- **Margin Bound** for IID setup (**generalization error controlled by ρ**):

$$\text{err}_P^{(\rho)}(f) \leq \text{err}_{\hat{P}}^{(\rho)}(f) + \frac{2k^2}{\rho} \mathfrak{R}_{n, P}(\Pi_1 \mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \quad (16)$$

Step II: Margin Disparity Discrepancy (MDD)⁷

- **Margin Disparity:** $\epsilon_D^{(\rho)}(f', f) \triangleq \mathbb{E}_{\mathbf{x} \sim D_X} [\Phi_\rho(\rho_{f'}(\mathbf{x}, h_f(\mathbf{x})))]$.
- We further define the margin version of Disparity Discrepancy (DD):

Definition (Margin Disparity Discrepancy (MDD))

Given a hypothesis space \mathcal{F} and a *specific hypothesis* $f \in \mathcal{F}$, the Margin Disparity Discrepancy (MDD) induced by $f' \in \mathcal{F}$ and its empirical version are defined by

$$\begin{aligned} d_{f, \mathcal{F}}^{(\rho)}(P, Q) &\triangleq \sup_{f' \in \mathcal{F}} \left(\epsilon_Q^{(\rho)}(f', f) - \epsilon_P^{(\rho)}(f', f) \right), \\ d_{f, \mathcal{F}}^{(\rho)}(\hat{P}, \hat{Q}) &\triangleq \sup_{f' \in \mathcal{F}} \left(\epsilon_{\hat{Q}}^{(\rho)}(f', f) - \epsilon_{\hat{P}}^{(\rho)}(f', f) \right). \end{aligned} \tag{17}$$

MDD satisfies $d_{f, \mathcal{F}}^{(\rho)}(P, P) = 0$ as well as **nonnegativity** and **subadditivity**.

⁷ Zhang & Long. *Bridging Theory and Algorithm for Domain Adaptation*. ICML 2019.

Margin Theory for Domain Adaptation

Theorem (Generalization Bound with Rademacher Complexity)

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be a hypothesis set with label set $\mathcal{Y} = \{1, \dots, k\}$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the corresponding \mathcal{Y} -valued labeling function class. Fix $\rho > 0$. For all $\delta > 0$, with probability $1 - 3\delta$ the following inequality holds for all hypothesis $f \in \mathcal{F}$:

$$\begin{aligned} \epsilon_Q(f) \leq & \epsilon_{\widehat{P}}^{(\rho)}(f) + d_{f, \mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} \\ & + \frac{2k^2}{\rho} \mathfrak{R}_{n, P}(\Pi_1 \mathcal{F}) + \frac{k}{\rho} \mathfrak{R}_{n, P}(\Pi_{\mathcal{H}} \mathcal{F}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\ & + \frac{k}{\rho} \mathfrak{R}_{m, Q}(\Pi_{\mathcal{H}} \mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \end{aligned} \quad (18)$$

An expected observation is that the **generalization risk is controlled by ρ** .

Margin Theory for Domain Adaptation

Theorem (Generalization Bound with Covering Numbers)

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be a hypothesis set with label set $\mathcal{Y} = \{1, \dots, k\}$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the corresponding \mathcal{Y} -valued labeling function class. Suppose $\Pi_1 \mathcal{F}$ is bounded in \mathcal{L}_2 by L . Fix $\rho > 0$. For all $\delta > 0$, with probability $1 - 3\delta$ the following inequality holds for all hypothesis $f \in \mathcal{F}$:

$$\begin{aligned} \epsilon_Q(f) \leq & \epsilon_{\hat{P}}^{(\rho)}(f) + d_{f, \mathcal{F}}^{(\rho)}(\hat{P}, \hat{Q}) + \epsilon_{ideal} + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\ & + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \frac{16k^2\sqrt{k}}{\rho} \inf_{\epsilon \geq 0} \left\{ \epsilon + 3\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \right. \\ & \left. \left(\int_{\epsilon}^L \sqrt{\log \mathcal{N}_2(\tau, \Pi_1 \mathcal{F})} d\tau + L \int_{\epsilon/L}^1 \sqrt{\log \mathcal{N}_2(\tau, \Pi_1 \mathcal{H})} d\tau \right) \right\}. \end{aligned} \quad (19)$$

The margin bound for OOD has **same order** with the margin bound for IID.

Hypothesis Adversarial Learning⁸

Minimax domain adaptation implied directly through the margin theory

$$\min_{f, \psi} \epsilon_{\psi(\hat{P})}^{(\rho)}(f) + \left(\epsilon_{\psi(\hat{Q})}^{(\rho)}(f^*, f) - \epsilon_{\psi(\hat{P})}^{(\rho)}(f^*, f) \right) \quad (20)$$
$$f^* = \max_{f'} \left(\epsilon_{\psi(\hat{Q})}^{(\rho)}(f', f) - \epsilon_{\psi(\hat{P})}^{(\rho)}(f', f) \right)$$

Theory

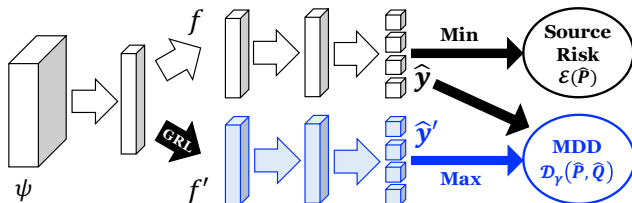
Bridge the Gap

Algorithm

1. Multiclass learning with scoring functions
2. Tight bound with only one hypothesis space
3. Informative bound with computable margin

⁸Zhang & Long. Bridging Theory and Algorithm for Domain Adaptation. ICML 2019.

Hypothesis Adversarial Learning



$$\begin{aligned}\mathcal{E}(\hat{P}) &= -\mathbb{E}_{(\mathbf{x}^s, y^s) \sim \hat{P}} \log[\sigma_{y^s}(f(\psi(\mathbf{x}^s)))] \\ \mathcal{D}_\gamma(\hat{P}, \hat{Q}) &= \mathbb{E}_{\mathbf{x}^t \sim \hat{Q}} \log[1 - \sigma_{h_f(\psi(\mathbf{x}^t))}(f'(\psi(\mathbf{x}^t)))] \\ &\quad + \gamma \mathbb{E}_{\mathbf{x}^s \sim \hat{P}} \log[\sigma_{h_f(\psi(\mathbf{x}^s))}(f'(\psi(\mathbf{x}^s)))]\end{aligned}\tag{21}$$

Theorem (Margin Implementation)

(Informal) Assuming that there is no restriction on the choice of f' and $\gamma > 1$, the global minimum of $\mathcal{D}_\gamma(P, Q)$ is $P = Q$. The value of $\sigma_{h_f}(f'(\cdot))$ at equilibrium is $\gamma/(1 + \gamma)$ and the corresponding margin of f' is $\rho = \log \gamma$.

Outline

1 Domain Adaptation

2 Theory and Algorithms

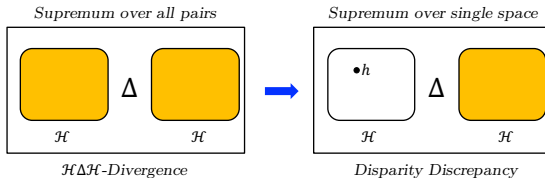
- Classic Theory
- Margin Theory
- Localization Theory

3 Open Library

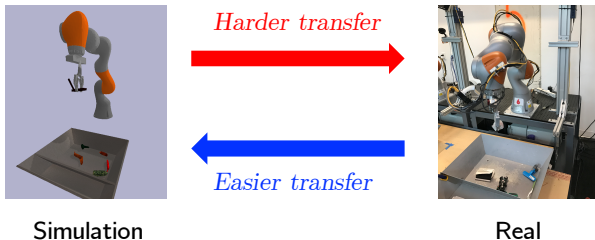
- TLLib: Transfer Learning Library

Theory vs. Practice

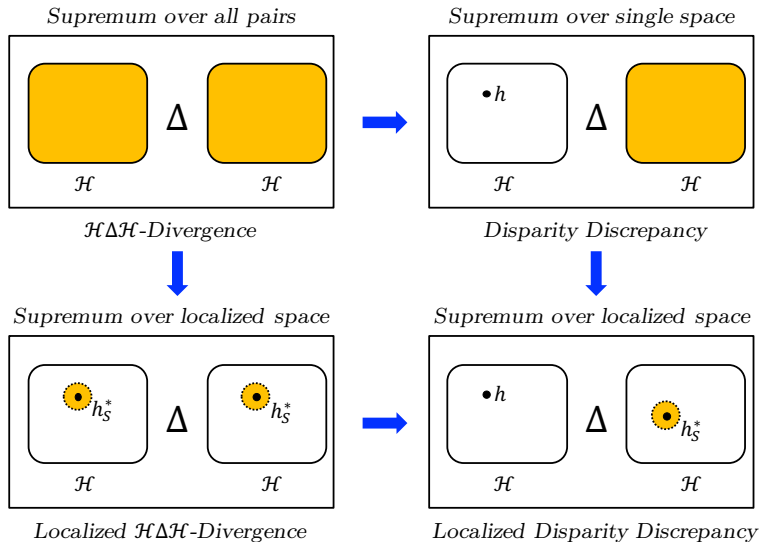
- Previous discrepancies are supremum over **whole hypothesis space** — will include bad hypotheses that make the bound **excessively large**.



- A common observation is that difficulty of transfer is **asymmetric** — Previous bounds will **remain unchanged** after switching P and Q .



Localization for Discrepancies



Step III: Localized Discrepancies

Definition (Localized Hypothesis Space)

For any distributions P and Q on $\mathcal{X} \times \mathcal{Y}$, any hypothesis space \mathcal{H} and any $r \geq 0$, the **localized hypothesis space** \mathcal{H}_r is defined as

$$\mathcal{H}_r = \{h \in \mathcal{H} | \mathbb{E}_P L(h(\mathbf{x}), y) \leq r\}. \quad (22)$$

Definition (Localized $\mathcal{H}\Delta\mathcal{H}$ -Discrepancy (LHH))

The **localized $\mathcal{H}\Delta\mathcal{H}$ -discrepancy** from P to Q is defined as

$$d_{\mathcal{H}_r\Delta\mathcal{H}_r}(P, Q) = \sup_{h, h' \in \mathcal{H}_r} (\mathbb{E}_Q L(h', h) - \mathbb{E}_P L(h', h)). \quad (23)$$

Definition (Localized Disparity Discrepancy (LDD))

For $h \in \mathcal{H}$, the **localized disparity discrepancy** from P to Q is defined as

$$d_{h, \mathcal{H}_r}(P, Q) = \sup_{h' \in \mathcal{H}_r} (\mathbb{E}_Q L(h', h) - \mathbb{E}_P L(h', h)). \quad (24)$$

Localization Theory for Domain Adaptation⁹

Recall the generalization bound induced by previous discrepancies:

$$\epsilon_Q(h) \leq \epsilon_{\hat{P}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q}) + \epsilon_{ideal} + O\left(\sqrt{\frac{d \log n}{n}} + \sqrt{\frac{d \log m}{m}}\right)$$

Theorem (Generalization Bound with Localized $\mathcal{H}\Delta\mathcal{H}$ -Discrepancy)

Set fixed $r > \lambda$. Let \hat{h} be the solution of the source error minimization. Then with probability no less than $1 - \delta$, we have

$$\begin{aligned} \text{err}_Q(\hat{h}) &\leq \text{err}_{\hat{P}}(\hat{h}) + d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\hat{P}, \hat{Q}) + \lambda + O\left(\frac{d \log n}{n} + \frac{d \log m}{m}\right) \\ &\quad + O\left(\sqrt{\frac{2rd \log n}{n}} + \sqrt{\frac{(d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\hat{P}, \hat{Q}) + 2r)d \log m}{m}}\right). \end{aligned} \quad (25)$$

To make domain adaptation feasible, we require $d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\hat{P}, \hat{Q}) + r \ll 1$.

⁹Zhang & Long. On Localized Discrepancy for Domain Adaptation. Preprint 2021.

Localization Theory for Domain Adaptation¹⁰

Recall that Disparity Discrepancy is **tighter** than $\mathcal{H}\Delta\mathcal{H}$ -Discrepancy:

$$\min_{\bar{h} \in \mathcal{H}} \{ \text{err}_{\hat{P}}(\bar{h}) + d_{\bar{h}, \mathcal{H}_r}(\hat{P}, \hat{Q}) \} \leq \min_{\hat{h} \in \mathcal{H}} \text{err}_{\hat{P}}(\hat{h}) + d_{\mathcal{H}_r \Delta \mathcal{H}_r}(\hat{P}, \hat{Q}) \quad (26)$$

Theorem (Generalization bound with localized disparity discrepancy)

Set fixed $r > \lambda$. Let \bar{h} be the solution of above left objective function. Then with probability no less than $1 - \delta$, we have

$$\begin{aligned} \text{err}_Q(\bar{h}) &\leq \text{err}_{\hat{P}}(\bar{h}) + d_{\bar{h}, \mathcal{H}_r}(\hat{P}, \hat{Q}) + \lambda + O\left(\frac{d \log n}{n} + \frac{d \log m}{m}\right) \\ &\quad + O\left(\sqrt{\frac{(\text{err}_{\hat{P}}(\bar{h}) + r)d \log n}{n}} + \sqrt{\frac{(\text{err}_{\hat{P}}(\bar{h}) + d_{\bar{h}, \mathcal{H}_r}(\hat{P}, \hat{Q}) + r)d \log m}{m}}\right). \end{aligned}$$

¹⁰Long et al. On Localized Discrepancy for Domain Adaptation. Preprint 2021.

Outline

1 Domain Adaptation

2 Theory and Algorithms

- Classic Theory
- Margin Theory
- Localization Theory

3 Open Library

- TLLib: Transfer Learning Library

Transfer Learning Library

📁 [thuml](#) / [Transfer-Learning-Library](#) Public

Transfer Learning Library for Domain Adaptation, Task Adaptation, and Domain Generalization

🔗 transfer.thuml.ai

📜 MIT license

★ 2k stars 🍴 396 forks



Updates

2022.9

We support installing *TLlib* via `pip`, which is experimental currently.

```
pip install -i https://test.pypi.org/simple/ tllib==0.4
```

2022.8

We release `v0.4` of *TLlib*. Previous versions of *TLlib* can be found [here](#). In `v0.4`, we add implementations of the following methods:

- Domain Adaptation for Object Detection [\[Code\]](#) [\[API\]](#)
- Pre-trained Model Selection [\[Code\]](#) [\[API\]](#)
- Semi-supervised Learning for Classification [\[Code\]](#) [\[API\]](#)

Design Patterns

Reproducible

Stable

Extendible

Ease of Use

TorchVision

Documentation

Docs

Examples

- ❑ Training codes
- ❑ Hyperparameters
- ❑

Benchmarks

- ❑ Various setups
- ❑ Reproducible
- ❑

Tutorials

- ❑ More data formats
- ❑ More model backbones
- ❑

Core

Adaptation

- ❑ DAN
- ❑ DANN
- ❑ MDD
- ❑ CDAN
- ❑

Module

- ❑ Discriminator
- ❑ GradRevLayer
- ❑ Kernel
- ❑

Backbone

- ❑ ResNet
- ❑ VGG
- ❑ Inception
- ❑

Dataset

- ❑ Office-31
- ❑ Office-Home
- ❑ VisDA-2017
- ❑ DomainNet
- ❑

Utils

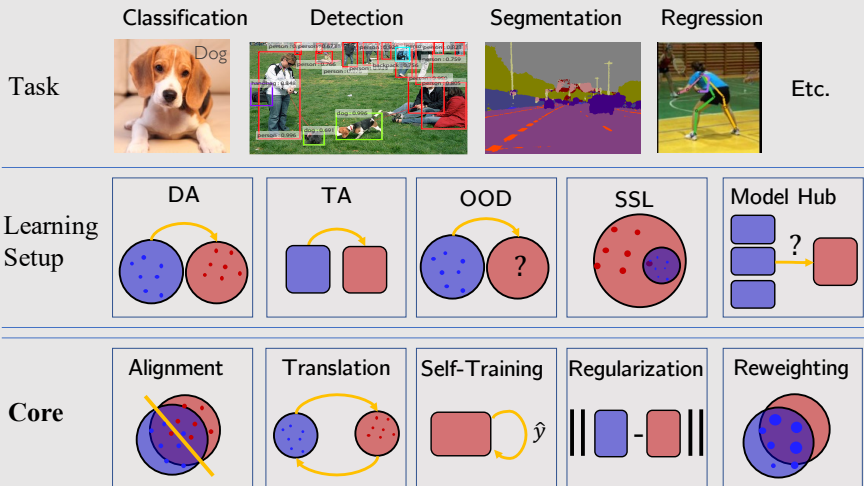
Platform



.....

Code: <https://github.com/thuml/Transfer-Learning-Library>















Learning Settings



<https://github.com/thuml/A-Roadmap-for-Transfer-Learning>

Benchmark Datasets

- **Vision:** Office-31, Office-Home, DomainNet, ImageCLEF, PACS, ...
- **Language:** Multi-Domain Sentiment Dataset, ...
- **WILDS:** Diverse domains <http://ai.stanford.edu/blog/wilds>
 - Still not work due to **large #domains**, **natural shifts**, and **harder tasks**

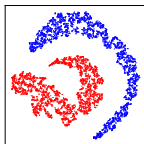
	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	lWldCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	<pre>import numpy as np ... norm=np.____</pre>
Test example						As a Christian, I will not be patronizing any of those businesses.			I "loved" my French press, it's so perfect and came with all this fun stuff!	<pre>import subprocess as sp p=sp.Popen() stdout=p.____</pre>
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

Regression on Visual Domains

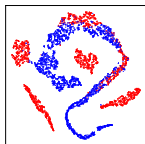
- Domain adaptation regression tasks on **dSprite** (simulation-to-real).

Method	$C \rightarrow N$	$C \rightarrow S$	$N \rightarrow C$	$N \rightarrow S$	$S \rightarrow C$	$S \rightarrow N$	Avg
ResNet-18	0.94	0.90	0.16	0.65	0.08	0.26	0.498
TCA	0.94	0.87	0.19	0.66	0.10	0.23	0.498
DAN	0.70	0.77	0.12	0.50	0.06	0.11	0.377
DANN	0.47	0.46	0.16	0.65	0.05	0.10	0.315
MCD	0.81	0.81	0.17	0.65	0.07	0.19	0.450
RSD	0.32	0.35	0.16	0.57	0.08	0.09	0.262
RSD+BMP	0.31	0.31	0.12	0.53	0.07	0.08	0.237
DD	0.09	0.27	0.07	0.21	0.12	0.11	0.145

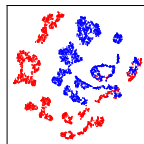
- To our knowledge, this is the **first** successful regression algorithm!



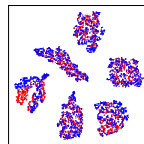
(a) Src Only



(b) DANN



(c) RSD







(d) DD

Classification on Visual Domains

- Domain adaptation classification tasks on **DomainNet** and **ImageNet**.

Methods	c→p	c→r	c→s	p→c	p→r	p→s	r→c	r→p	r→s	s→c	s→p	s→r	Avg
ResNet101	32.7	50.6	39.4	41.1	56.8	35.0	48.6	48.8	36.1	49.0	34.8	46.1	43.3
DAN	38.8	55.2	43.9	45.9	59.0	40.8	50.8	49.8	38.9	56.1	45.9	55.5	48.4
DANN	37.9	54.3	44.4	41.7	55.6	36.8	50.7	50.8	40.1	55.0	45.0	54.5	47.2
JAN	40.5	56.7	45.1	47.2	59.9	43.0	54.2	52.6	41.9	56.6	46.2	55.5	50.0
CDAN	40.4	56.8	46.1	45.1	58.4	40.5	55.6	53.6	43.0	57.2	46.4	55.7	49.9
MDD	42.9	59.5	47.5	48.6	59.4	42.6	58.3	53.7	46.2	58.7	46.5	57.7	51.8

Task	IN→INR (ResNet50)	IN→INS (ig_resnext101_32x8d)
Source Only	35.6	54.9
DAN	39.8	55.7
DANN	52.7	56.5
JAN	41.7	55.7
CDAN	53.9	58.2
MDD	56.2	62.4

	clipart	real	sketch	painting
Image				
Category	Watermelon	Cow	Tree	Bird

(e) DomainNet

	ImageNet	ImageNet-R	ImageNet-Sketch
Image			
Category	Daisy	Guinea Pig, Cavia Cobaya	Promontory, Headland, Foreland

(f) ImageNet

Classification on Language Domains

- Domain adaptation classification tasks on **Text Emotion** datasets.

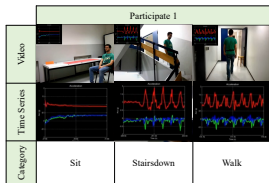
Methods	g→s	g→i	g→e	s→i	s→e	i→e	avg
Source Only	45.4	47.0	58.2	31.5	33.7	50.6	44.4
DANN	38.3	44.3	57.3	32.5	39.4	54.5	44.4
DAN	34.6	42.7	57.1	34.5	37.7	49.6	42.7
JAN	32.6	43.1	57.6	34.2	39.4	50.3	42.9
CDAN	39.7	46.2	57.2	34.6	35.3	52.1	44.2
MDD	46.3	47.3	72.8	38.8	48.1	65.8	53.2
LDD	49.4	49.0	75.9	43.2	50.7	68.4	56.1

Dataset	emotions	examples
GoEmotions	admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise	OMG, yep!!! That is the final answer. Thank you so much! Let me give you a hint: THEY PLAY IN BOSTON!!! I think this is my favourite one ever. I think that question has a very complicated answer I'm not even sure what it is, why do people hate it
SemEval-2018	anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust	It appears my fire alarm disapproves of my cooking style Heyyyy warriors!!!! #panicattacks
Emotions-stimulus	sadness, joy, anger, fear, surprise, disgust	There was a hint of exasperation in his voice. He could see Harry's puzzlement.
ISEAR	joy, fear, anger, sadness, disgust, shame, guilt	I am happy when I get good results in the field of academics or athletics. My cat died from an illness. It had been with us for 7 years.

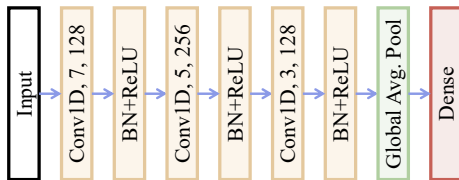
Classification on Time-Series Domains

- Domain adaptation classification tasks on **Human Activity** datasets.

Methods	1→3	3→5	4→5	1→6	4→6	5→6	3→8	5→8	Average
FCN	90.3	82.9	74.9	83.2	61.9	58.8	81.4	91.8	78.2
DAN	91.2	93.7	89.7	89.7	78.0	80.2	86.8	93.3	87.8
DANN	92.5	95.9	93.1	91.8	78.8	91.6	95.2	96.6	91.9
VRADA	81.3	82.3	71.6	74.9	62.7	60.0	82.2	87.5	75.3
CDAN	93.9	96.8	95.2	92.9	83.6	91.6	95.6	97.6	93.4
R-DANN	85.1	85.4	70.4	81.7	64.6	54.4	82.8	82.5	75.9
CoDATS	93.2	95.6	94.2	90.5	93.7	90.7	93.4	97.1	93.5
CoDATS+WS	90.8	94.3	94.7	90.8	85.3	91.7	94.3	95.8	92.2
MDD	95.5	97.9	96.9	93.2	84.6	91.7	96.9	97.9	94.3

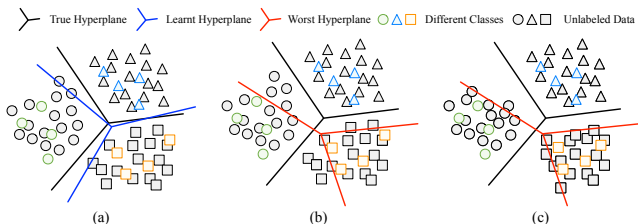
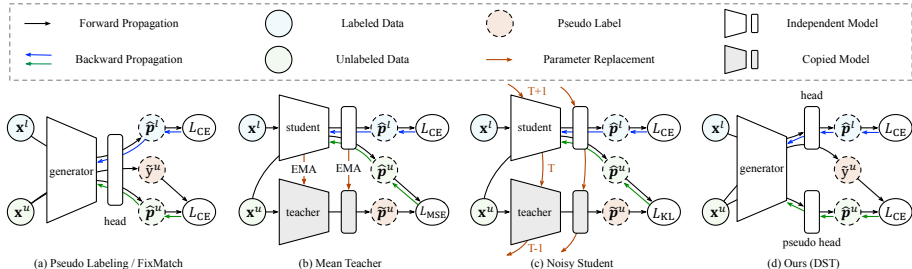


(g) HHAR



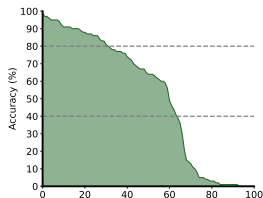
(h) FFN

Distribution Shift in Semi-Supervised Learning

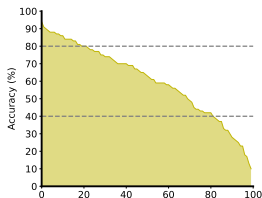


$$\min_{\psi, h, h_{\text{pseudo}}} L_{\mathcal{L}}(\psi, h) + L_{\mathcal{U}}(\psi, h_{\text{pseudo}}, \hat{f}_{\psi, h}) + \max_{h'} (L_{\mathcal{U}}(\psi, h', \hat{f}_{\psi, h}) - L_{\mathcal{L}}(\psi, h')).$$

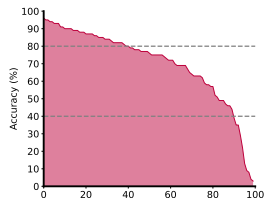
Distribution Shift in Semi-Supervised Learning



(i) FixMatch



(j) DST w/o worst



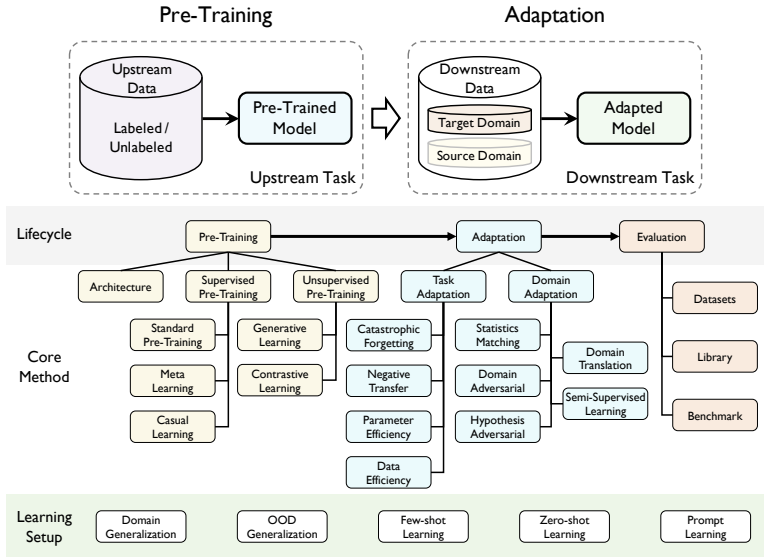
(k) DST

Figure: Top-1 accuracy of each category on *CIFAR-100* (supervised pre-trained).

FixMatch	86.3	84.6	53.1	41.3	48.6	25.2	52.3	93.2	83.7	46.4	37.1	59.3
DST (FixMatch)	89.6	94.9	70.4	48.1	53.5	43.2	68.7	94.8	89.8	71.0	58.5	71.1
FixMatch	83.1	82.2	51.4	39.2	43.9	30.1	36.8	94.3	65.7	48.6	36.8	55.6
DST (FixMatch)	90.1	95.0	68.2	46.8	54.2	47.7	53.6	95.6	75.4	72.0	57.1	68.7

Chen et al. Debiased Self-Training for Semi-Supervised Learning. NeurIPS 2022 (Oral).

Transferability in Deep Learning: A Small Book¹¹



¹¹<https://arxiv.org/abs/2201.05867>

Machine Learning Group @ THSS



Mingsheng Long
(龙明盛)

Tsinghua University
mingsheng@tsinghua.edu.cn



Jianmin Wang
(王建民)

Tsinghua University
jimwang@tsinghua.edu.cn



Michael I. Jordan
(迈克尔·欧文·乔丹)

UC Berkeley
jordan@cs.berkeley.edu



Yuchen Zhang
(张育宸)



Zhangjie Cao
(曹张杰)



Han Zhu
(朱晗)



Yue Cao
(曹越)



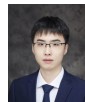
Kaichao You
(游凯超)



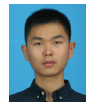
Janguang Jiang
(江俊广)



Ximei Wang
(王希梅)



Xinyang Chen
(陈新阳)



Yang Shu
(树扬)



长按关注，获取最新资讯