Transfer Learning:
Theories, Algorithms, and Open Library

Mingsheng Long

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

mingsheng@tsinghua.edu.cn
http://ise.thss.tsinghua.edu.cn/~mlong
Workshop on Federated and Transfer Learning, FTL-IJCAI'21

Supervised Learning

Learner: $f : \boldsymbol{x} \rightarrow y$ **Distribution:** $(\boldsymbol{x}, y) \sim P(\boldsymbol{x}, y)$

fish

bird

mammal

tree

flower

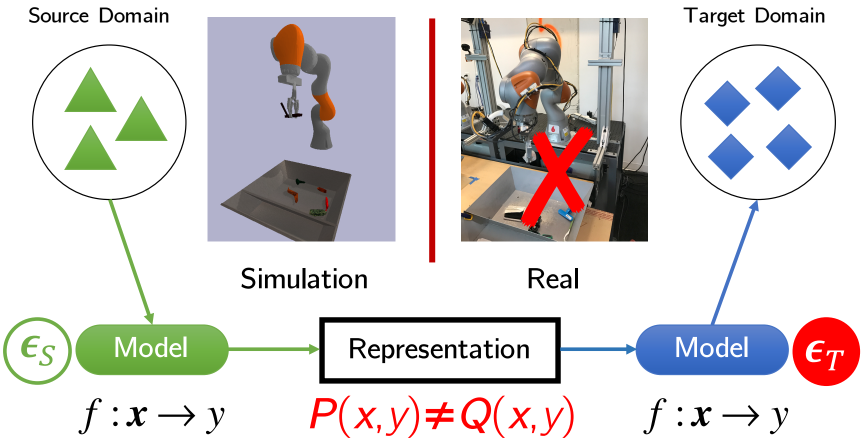
......

IID Setup

Error Bound: $\epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{complexity}}{n}}$

Transfer Learning

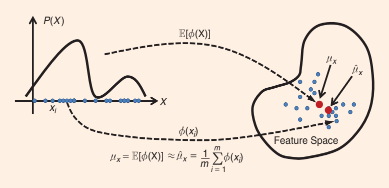
- Machine learning across domains of different distributions $P \neq Q$
  - **OOD: Out-of-Distribution** (from IID to OOD)
- How to bound generalization error on target domain for OOD setup?

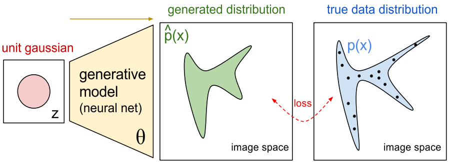


Representative Approaches to Transfer Learning

Learning to **match distributions** across **OOD** domains *s.t.* $P \approx Q$

- Covariate shift: $P(\mathbf{X}) \neq Q(\mathbf{X})$ (mainstream work of this setup)
- Prior shift: $P(\mathbf{Y}) \neq Q(\mathbf{Y})$ (challenging, current hotspot)
- Conditional shift: $P(Y|\mathbf{X}) \neq Q(Y|\mathbf{X})$ (challenging, future research)



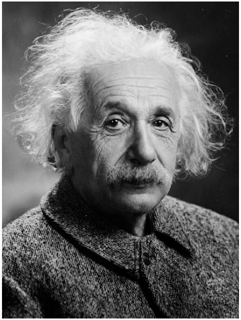


Kernel Embedding **Adversarial Learning**

Generally, no theoretical guarantees!

Song et al. Kernel Embeddings of Conditional Distributions. ***IEEE***, *2013.*
Goodfellow et al. Generative Adversarial Networks. ***NIPS*** *2014.*

Principal Problem: Bridging Theory and Algorithm



Everything should be made as simple as possible, but no simpler.

—Albert Einstein

There is nothing more practical than a good theory.

—Vladimir Vapnik

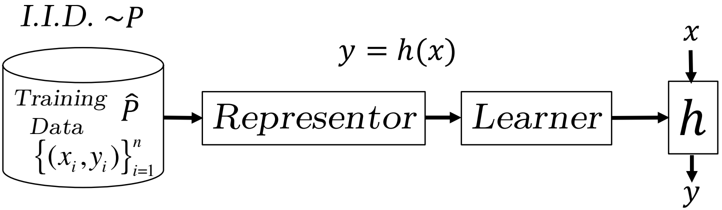
Outline

1. Transfer Learning

2. **Theories and Algorithms**
   - Classic Theory
   - Margin Theory
   - Localization Theory

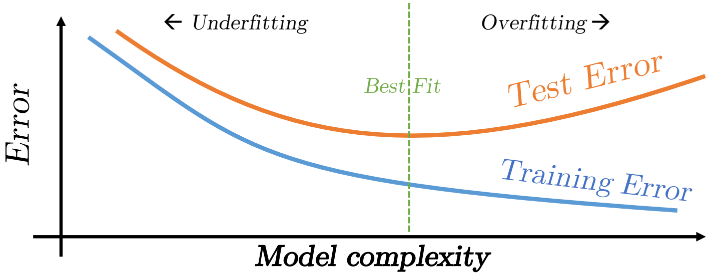
3. Open Library
   - Transfer-Learning-Library

Statistical Learning



- Formally analyzing the classification problem with **01-loss** $[\cdot \neq \cdot]$.
- Training error: $\epsilon_{\widehat{P}}(h) = \frac{1}{n}\sum_{i=1}^{n}[h(\mathbf{x}_i) \neq y_i] = \mathbb{E}_{(\mathbf{x},y)\sim\widehat{P}}[h(\mathbf{x}) \neq y]$.
- Test error: $\epsilon_P(h) = \mathbb{E}_{(\mathbf{x},y)\sim P}[h(\mathbf{x}) \neq y]$.
- Training error is an *unbiased* estimation of test error.
- Principal problem: Can we control $\epsilon_P(h)$ with observable $\epsilon_{\widehat{P}}(h)$?

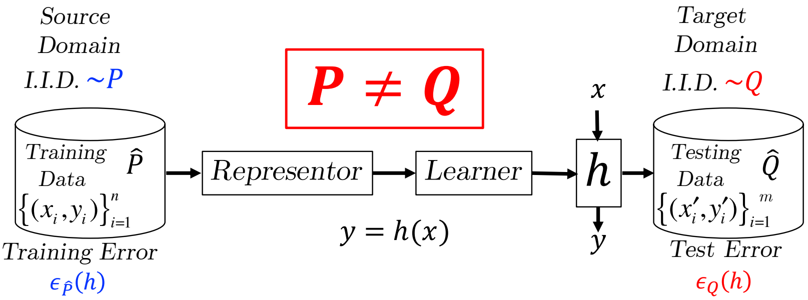
Statistical Learning Theory



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

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

Transfer Learning



- Only have labeled data sampled from a different source domain P.
- And unlabeled data sampled from a target domain Q. $\epsilon_{\hat{Q}}(h)$ is not observable!
- Principal problem: Can we control target error $\epsilon_Q(h)$?
- **Disparity on** D: $\epsilon_D(h_1, h_2) = \mathbb{E}_{(\mathbf{x},y)\sim D}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$.
- Why use it? Computation of disparity does not require (target) label!

Relating Target Risk to Source Risk

Theorem (Bound with Disparity)

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

$$\epsilon_Q(h) \leqslant \epsilon_P(h) + \left[\epsilon_P(h^*) + \epsilon_Q(h^*)\right] + \left|\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)\right| \tag{1}$$

Proof.

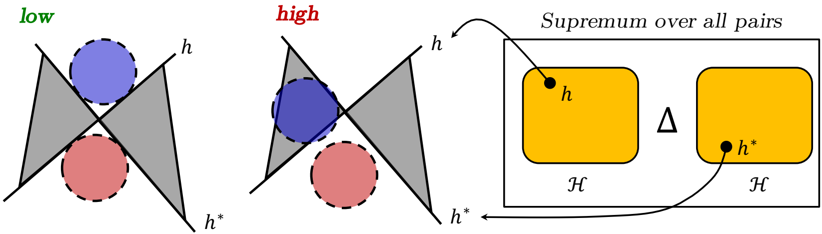
Simply using the *triangle inequalities* of the 01-loss, we have

$$\begin{aligned}
\epsilon_Q(h) &\leqslant \epsilon_Q(h^*) + \epsilon_Q(h, h^*) \\
&= \epsilon_Q(h^*) + \epsilon_P(h, h^*) + \epsilon_Q(h, h^*) - \epsilon_P(h, h^*) \\
&\leqslant \epsilon_Q(h^*) + \epsilon_P(h, h^*) + \left|\epsilon_Q(h, h^*) - \epsilon_P(h, h^*)\right| \\
&\leqslant \epsilon_P(h) + \left[\epsilon_P(h^*) + \epsilon_Q(h^*)\right] + \left|\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)\right|
\end{aligned} \tag{2}$$

□

$\mathcal{H}\Delta\mathcal{H}$-Divergence[1]

- **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.

[1] *Ben-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

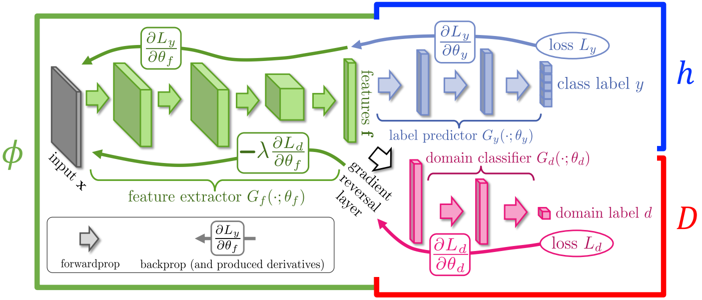
$$\epsilon_Q(h) \leq \epsilon_{\hat{P}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} + O\left(\sqrt{\frac{d \log n}{n}} + \sqrt{\frac{d \log m}{m}}\right) \quad (3)$$

- 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

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

- This bound can be estimated by training a domain discriminator $D(\mathbf{x})$.
- It can also be approximated by the Integral Probability Metric (IPM).

Domain Adversarial Neural Network (DANN)[2]



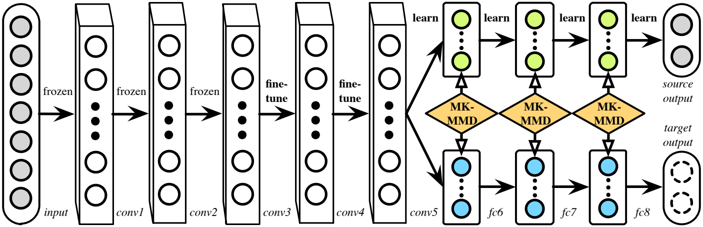
Adversarial domain adaptation: learn ϕ to minimize $d_{\mathcal{H}\Delta\mathcal{H}}(\phi(P), \phi(Q))$.

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

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

[2] *Ganin 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

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

$$\min_{\theta \in \Theta} \max_{k \in \mathcal{K}} \frac{1}{n_a} \sum_{i=1}^{n_a} L\left(\theta\left(\mathbf{x}_i^a\right), y_i^a\right) + \lambda \sum_{\ell=l_1}^{l_2} d_k^2\left(\widehat{P}_\ell, \widehat{Q}_\ell\right) \tag{7}$$

Works better than f-Divergences when domains are less overlapping

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

Outline

1. Transfer Learning
2. **Theories and Algorithms**
   - Classic Theory
   - Margin Theory
   - Localization Theory
3. Open Library
   - Transfer-Learning-Library

Theory and Practice: Gap Exists for Decade

- **Theory vs. Practice:**
- Binary Classification vs. Multiclass Classification.
- Discrete Classifier vs. Classifier with Scoring Function.
- $d_{\mathcal{H}\Delta\mathcal{H}}$ does not need label vs. $d_{\mathcal{H}\Delta\mathcal{H}}$ is hard to compute and optimize.
- **Principal problem: How to bridge 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

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

Theorem (Bound with Disparity Discrepancy)

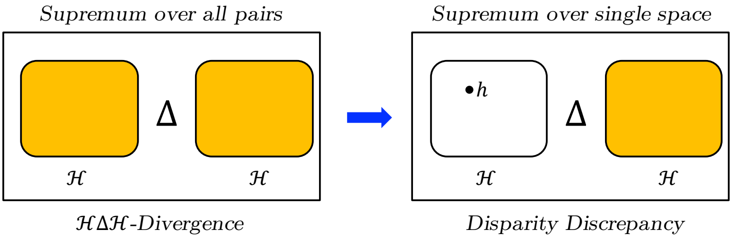
For any $\delta > 0$ and binary classifier $h \in \mathcal{H}$, with probability $1 - 3\delta$, we have

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

[4] *Zhang & 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.

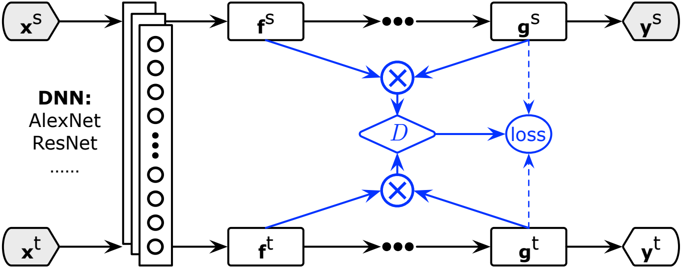


- DD can be estimated by conditional domain discriminator $D(\mathbf{x}, h(\mathbf{x}))$.

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

- It can also be approximated by the Integral Probability Metric (IPM).

Conditional Domain Adversarial Network (CDAN)[5]



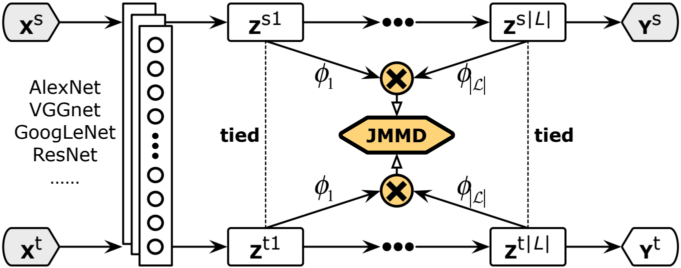
Conditional adversarial domain adaptation: minimize $d_{h,\mathcal{H}}(\phi(P), \phi(Q))$.

$$\begin{aligned} &\min_{G} \ \mathcal{E}(G) - \lambda \mathcal{E}(D, G) \\ &\min_{D} \ \mathcal{E}(D, G), \end{aligned} \tag{11}$$

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

[5] *Long et al. Conditional Adversarial Domain Adaptation. NIPS 2018.*

Joint Adaptation Network (JAN)[6]



Joint distribution matching: cross-covariance of multiple random vectors

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

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

Works better than f-Divergences when domains are less overlapping

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

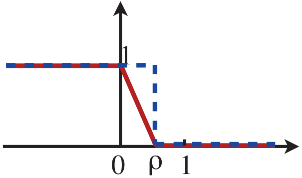
Multiclass Classification Formulation

- Scoring function: $f \in \mathcal{F} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$
- 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:

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

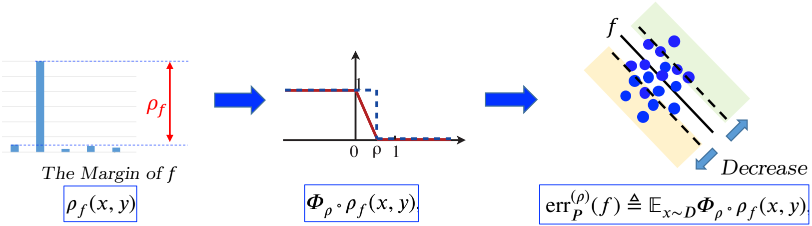
- Margin Loss:

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



Margin Theory

- Margin error: $\epsilon_D^{(\rho)}(f) = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[\Phi_\rho(\rho_f(\mathbf{x},y))\right]$
- 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

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

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

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

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

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

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 Transfer Learning

Theorem (Generalization Bound with Rademacher Complexity)

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be a hypothesis set with label set $\mathcal{Y} = \{1, \cdots, k\}$ 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 $f \in \mathcal{F}$:

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

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

Margin Theory for Transfer Learning

Theorem (Generalization Bound with Covering Numbers)

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set with label set $\mathcal{Y} = \{1, \cdots, k\}$ 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 $f \in \mathcal{F}$:

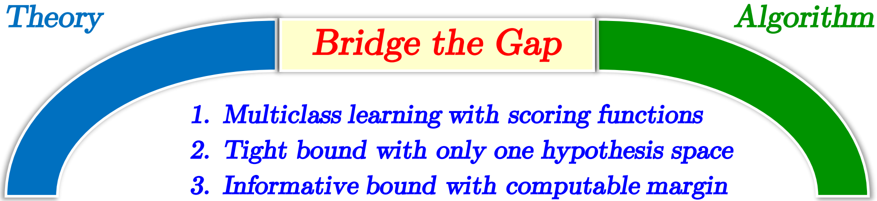
$$\begin{aligned}\epsilon_Q(f) \leq& \epsilon_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\ &+ \sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \frac{16k^2\sqrt{k}}{\rho} \inf_{\epsilon\geq 0}\Big\{\epsilon + 3\big(\frac{1}{\sqrt{n}}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

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

Margin Theory Implied Algorithm (MDD)[8]

Minimax domain adaptation implied directly through the margin theory

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



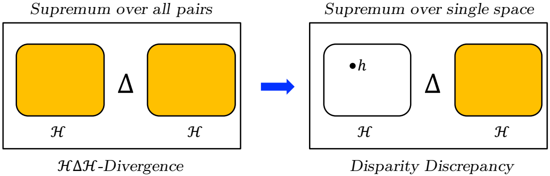
[8] *Zhang & Long. Bridging Theory and Algorithm for Domain Adaptation. ICML 2019.*

Outline

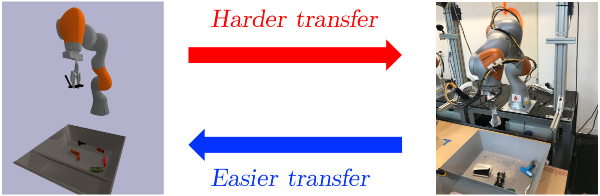
1. Transfer Learning
2. Theories and Algorithms
   - Classic Theory
   - Margin Theory
   - Localization Theory
3. Open Library
   - Transfer-Learning-Library

Theory and Practice: Final Gap to Close

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

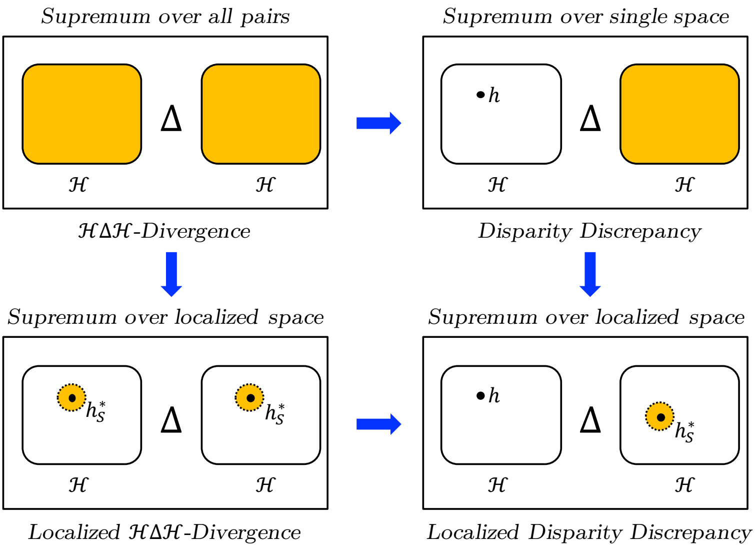


- A common observation is that difficulty of transfer is asymmetric — Previous bounds will remain unchanged after switching P and Q.



Simulation Real

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

$$\mathcal{H}_r = \{h \in \mathcal{H} | \mathbb{E}_P L(h(\mathbf{x}), y) \leq r\}. \tag{21}$$

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

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

$$d_{\mathcal{H}_r \Delta \mathcal{H}_r}(P, Q) = \sup_{h, h' \in \mathcal{H}_r} \left(\mathbb{E}_Q L(h', h) - \mathbb{E}_P L(h', h) \right). \tag{22}$$

Definition (Localized Disparity Discrepancy (LDD))

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

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

Localization Theory for Transfer Learning[9]

Recall the generalization bound induced by previous discrepancies:

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

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

$$\begin{aligned}\mathrm{err}_Q(\hat{h}) \leq \mathrm{err}_{\widehat{P}}(\hat{h}) &+ d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P}, \widehat{Q}) + \lambda + O(\frac{d \log n}{n} + \frac{d \log m}{m}) \\ &+ O\left(\sqrt{\frac{2rd \log n}{n}} + \sqrt{\frac{(d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P}, \widehat{Q}) + 2r)d \log m}{m}}\right).\end{aligned} \tag{24}$$

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

[9] *Zhang & Long. On Localized Discrepancy for Domain Adaptation. Preprint 2020.*

Localization Theory for Transfer Learning[10]

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

$$\min_{\bar{h}\in\mathcal{H}}\{\mathrm{err}_{\widehat{P}}(\bar{h}) + d_{\bar{h},\mathcal{H}_r}(\widehat{P},\widehat{Q})\} \le \min_{\hat{h}\in\mathcal{H}} \mathrm{err}_{\widehat{P}}(\hat{h}) + d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P},\widehat{Q}) \tag{25}$$

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

$$\mathrm{err}_Q(\hat{h}) \le \mathrm{err}_{\widehat{P}}(\bar{h}) + d_{\bar{h},\mathcal{H}_r}(\widehat{P},\widehat{Q}) + \lambda + O\left(\frac{d\log n}{n} + \frac{d\log m}{m}\right)$$
$$+ O\left(\sqrt{\frac{(\mathrm{err}_{\widehat{P}}(\bar{h}) + r)d\log n}{n}} + \sqrt{\frac{(\mathrm{err}_{\widehat{P}}(\bar{h}) + d_{\bar{h},\mathcal{H}_r}(\widehat{P},\widehat{Q}) + r)d\log m}{m}}\right). \tag{26}$$

[10] *Zhang & Long. On Localized Discrepancy for Domain Adaptation. Preprint 2020.*

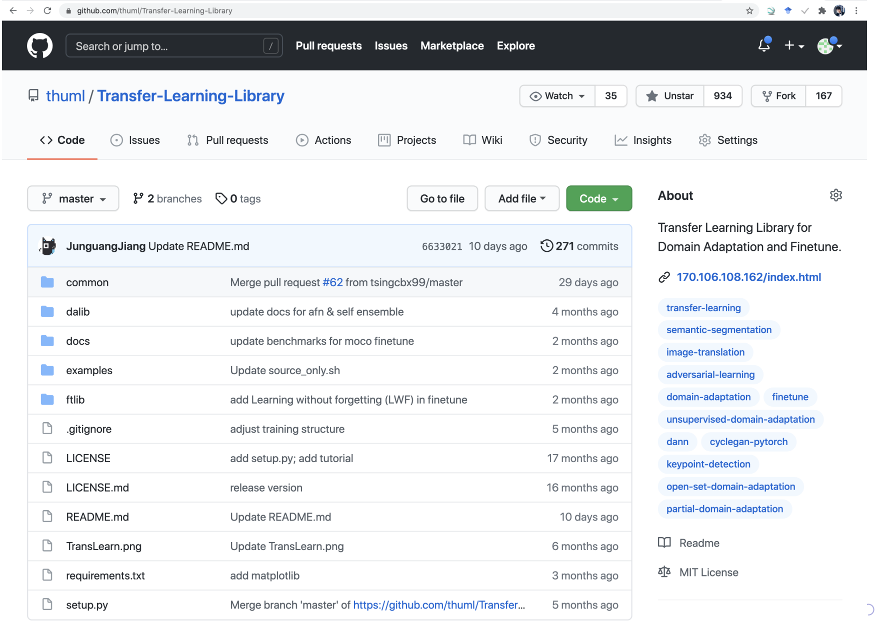
Outline

1 Transfer Learning

2 Theories and Algorithms
- Classic Theory
- Margin Theory
- Localization Theory

3 Open Library
- Transfer-Learning-Library

Transfer Learning Library



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

PyTorch

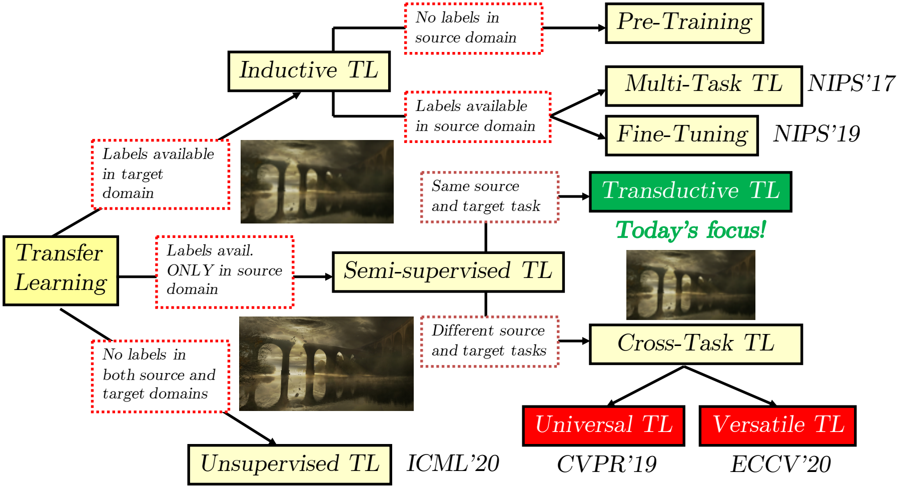
TorchVision

Facebook Open Source

......

Github: `https://github.com/thuml/Transfer-Learning-Library`

Standardized Implementations



This taxonomy was initiated by **Prof Q. Yang**, most setups are still open!

Reproducible Benchmarks

Table: Accuracy (%) on *Office-31* for Unsupervised Domain Adaptation

Method	Origin	Ours	Δacc	$A \to W$	$D \to W$	$W \to D$	$A \to D$	$D \to A$	$W \to A$
ResNet-50	76.1	**79.5**	3.4	75.8	95.5	99.0	79.3	63.6	63.8
DANN	82.2	**86.4**	4.2	91.7	97.9	100.0	82.9	72.8	73.3
DAN	80.4	**83.7**	3.3	84.2	98.4	100.0	87.3	66.9	65.2
JAN	84.3	**87.3**	3.0	93.7	98.4	100.0	89.4	71.2	71.0
CDAN	87.7	**88.7**	1.0	93.1	98.6	100.0	93.4	75.6	71.5
MCD	-	**85.9**	-	91.8	98.6	100.0	89.0	69.0	66.9
MDD	88.9	**89.2**	0.3	93.6	98.6	100.0	93.6	76.7	72.9

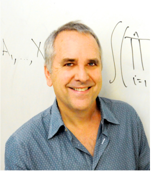
Table: Accuracy (%) on *Office-Home* for Unsupervised Domain Adaptation

Method	Origin	Ours	Δacc	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Pr-Ar	Pr-Cl	Pr-Rw	Rw-Ar	Rw-Cl	Rw-Pr
ResNet-50	46.1	**58.4**	12.3	41.1	65.9	73.7	53.1	60.1	63.3	52.2	36.7	71.8	64.8	42.6	75.2
DANN	57.6	**65.2**	7.6	53.8	62.6	74.0	55.8	67.3	67.3	55.8	55.1	77.9	71.1	60.7	81.1
DAN	56.3	**61.4**	5.1	45.6	67.7	73.9	57.7	63.8	66.0	54.9	40.0	74.5	66.2	49.1	77.9
JAN	58.3	**65.9**	7.6	50.8	71.9	76.5	60.6	68.3	68.7	60.5	49.6	76.9	71.0	55.9	80.5
CDAN	65.8	**68.8**	3.0	55.2	72.4	77.6	62.0	69.7	70.9	62.4	54.3	80.5	75.5	61.0	83.8
MCD	-	**67.8**	-	51.7	72.2	78.2	63.7	69.5	70.8	61.5	52.8	78.0	74.5	58.4	81.8
MDD	68.1	**69.6**	1.5	56.4	75.3	78.4	63.2	73.1	73.3	63.9	54.8	79.7	73.2	60.7	83.7

Machine Learning Group @ National Engineering Lab

Jianmin Wang
Professor
Tsinghua University
jimwang@tsinghua.edu.cn

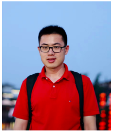
Mingsheng Long
Associate Professor
Tsinghua University
mingsheng@tsinghua.edu.cn



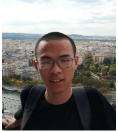
Michael I. Jordan
Professor
UC Berkeley
jordan@cs.berkeley.edu

Yuchen Zhang

Ximei Wang



Zhangjie Cao



Kaichao You

Junguang Jiang

Many thanks for your attention! Any questions?