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: $f : x \rightarrow y$
Distribution: $(x, y) \sim P(x, y)$

$IID$ Setup

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

fish
bird
mammal
tree
flower
......
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?

Source Domain: Simulation | Target Domain: Real

Model $\mathcal{E}_S$: $f: x \rightarrow y$

Representation

$P(x, y) \neq Q(x, y)$

Model $\mathcal{E}_T$: $f: x \rightarrow y$
### Representative Approaches to Transfer Learning

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

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

![Feature Space](image1.png)

**Kernel Embedding**

**Adversarial Learning**

**Generally, no theoretical guarantees!**


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: \(\epsilon_{\hat{P}}(h) = \frac{1}{n} \sum_{i=1}^{n} [h(x_i) \neq y_i] = \mathbb{E}_{(x,y) \sim \hat{P}}[h(x) \neq y]\).

Test error: \(\epsilon_P(h) = \mathbb{E}_{(x,y) \sim P}[h(x) \neq y]\).

Training error is an unbiased estimation of test error.

Principal problem: Can we control \(\epsilon_P(h)\) with observable \(\epsilon_{\hat{P}}(h)\)?
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:

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

Only have labeled data sampled from a different source domain \( P \).

And unlabeled data sampled from a target domain \( Q \). \( \epsilon_{\widehat{Q}}(h) \) is not observable!

Principal problem: Can we control target error \( \epsilon_{Q}(h) \)?

Disparity on \( D \): \( \epsilon_D(h_1, h_2) = \mathbb{E}_{(x,y) \sim D} [h_1(x) \neq h_2(x)] \).

Why use it? Computation of disparity does not require (target) label!
Relating Target Risk to Source Risk

Theorem (Bound with Disparity)

For classification tasks of transfer learning, 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:

\[
\epsilon_Q(h) \leq \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)| \tag{1}
\]

Proof.

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

\[
\epsilon_Q(h) \leq \epsilon_Q(h^*) + \epsilon_Q(h, h^*) \\
= \epsilon_Q(h^*) + \epsilon_P(h, h^*) + \epsilon_Q(h, h^*) - \epsilon_P(h, h^*) \\
\leq \epsilon_Q(h^*) + \epsilon_P(h, h^*) + |\epsilon_Q(h, h^*) - \epsilon_P(h, h^*)| \\
\leq \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)| \tag{2}
\]
**$\mathcal{H}\mathcal{D}\mathcal{H}$-Divergence**

- **Assumption:** Small ideal joint error $\epsilon_{\text{ideal}} = \epsilon_P(h^*) + \epsilon_Q(h^*)$.
- 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.

---

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

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

- 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

$$d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) \triangleq \sup_{h, h' \in \mathcal{H}} |\epsilon_P(h, h') - \epsilon_Q(h, h')|$$

$$= \sup_{\delta \in \mathcal{H}\Delta\mathcal{H}} |\mathbb{E}_P[\delta(x) \neq 0] - \mathbb{E}_Q[\delta(x) \neq 0]| \quad (4)$$

$$\leq \sup_{D \in \mathcal{H}_D} |\mathbb{E}_P[D(x) = 1] + \mathbb{E}_Q[D(x) = 0]|$$

- This bound can be estimated by training a domain discriminator $D(x)$.
- It can also be approximated by the Integral Probability Metric (IPM).
Domain Adversarial Neural Network (DANN)²

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

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

**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)\textsuperscript{3}

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

\begin{align}
    d_k^2 (P, Q) & \triangleq \left\| \mathbb{E}_P [\phi (x^s)] - \mathbb{E}_Q [\phi (x^t)] \right\|_H^2 \\
    \min_{\theta \in \Theta} \max_{k \in \mathcal{K}} \frac{1}{n_a} \sum_{i=1}^{n_a} L (\theta (x_i^a), y_i^a) + \lambda \sum_{\ell=1}^{l_2} d_k^2 (\hat{P}_\ell, \hat{Q}_\ell) 
\end{align}

Works better than $f$-Divergences when domains are less overlapping

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

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

(8)

**Theorem (Bound with Disparity Discrepancy)**

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

\[
\epsilon_Q(h) \leq \epsilon_{\hat{P}}(h) + d_{h,\mathcal{H}}(\hat{P}, \hat{Q}) + \epsilon_{\text{ideal}} + 2\mathfrak{R}_{n,P}(\mathcal{H} \Delta \mathcal{H})
\]

\[
+ 2\mathfrak{R}_{n,P}(\mathcal{H}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} + 2\mathfrak{R}_{m,Q}(\mathcal{H} \Delta \mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.
\]

(9)

Step I: Disparity Discrepancy (DD)

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

\[ \sup_{h', \in \mathcal{H}} (\varepsilon_P (h, h') - \varepsilon_Q (h, h')) \]
\[ = \sup_{h', \in \mathcal{H}} (\mathbb{E}_P [|h(x) - h'(x)| \neq 0] - \mathbb{E}_Q [|h(x) - h'(x)| \neq 0]) \]
\[ \leq \sup_{D \in \mathcal{H}_D} (\mathbb{E}_P [D(x, h(x)) = 1] + \mathbb{E}_Q [D(x, h(x)) = 0]) \]

- DD can be estimated by **conditional domain discriminator** $D(x, h(x))$.

It can also be approximated by the Integral Probability Metric (IPM).
Conditional adversarial domain adaptation: minimize $d_{h,\mathcal{H}}(\phi(P), \phi(Q))$.

\[
\min_G \mathcal{E}(G) - \lambda \mathcal{E}(D, G)
\]
\[
\min_D \mathcal{E}(D, G),
\]

\[
\mathcal{E}(D, G) = -\mathbb{E}_{x_i \sim D_s} \log [D(f_i^s \otimes g_i^s)] - \mathbb{E}_{x_j \sim D_t} \log [1 - D(f_j^t \otimes g_j^t)]
\]

---

Joint Adaptation Network (JAN)

Joint distribution matching: cross-covariance of multiple random vectors

\[ d^2_k (P, Q) \triangleq \left\| \mathbf{E}_P \left[ \bigotimes_{\ell=1}^m \phi_{\ell} (x^s_\ell) \right] - \mathbf{E}_Q \left[ \bigotimes_{\ell=1}^m \phi_{\ell} (x^t_\ell) \right] \right\|^2_{\mathcal{H}_k} \]  \hspace{1cm} (13)

\[
\min_{\theta \in \Theta} \max_{k \in \mathcal{K}} \frac{1}{n_a} \sum_{i=1}^{n_a} L(\theta (x^a_i), y^a_i) + \lambda d^2_k \left( \hat{P}_{\ell=1:L}, \hat{Q}_{\ell=1:L} \right) \]  \hspace{1cm} (14)

Works better than \( f \)-Divergences when domains are less overlapping

---

Multiclass Classification Formulation

- **Scoring function:** \( f \in \mathcal{F} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \)
- **Labeling function** induced by \( f \): \( h_f : x \mapsto \arg \max_{y \in \mathcal{Y}} f(x, y) \)
- **Labeling function class:** \( \mathcal{H} = \{ h_f | f \in \mathcal{F} \} \)
- **Margin** of a hypothesis \( f \):
  \[
  \rho_f(x, y) = \frac{1}{2}(f(x, y) - \max_{y' \neq y} f(x, y'))
  \]
- **Margin Loss:**
  \[
  \Phi_\rho(x) = \begin{cases} 
  0 & \rho \leq x \\
  1 - x/\rho & 0 \leq x \leq \rho \\
  1 & x \leq 0 \end{cases}
  \]
Margin Theory

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

\[
\Phi_{\rho} \cdot \rho_f(x, y).
\]

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

\[
\Pi_1 \mathcal{F} = \{ x \mapsto f(x, y) | y \in \mathcal{Y}, f \in \mathcal{F} \}. \tag{15}
\]

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

\[
\text{err}_P^{(\rho)} (f) \leq \text{err}_P^{(\rho)} (f) + \frac{2k^2}{\rho} \mathcal{R}_{n,P}(\Pi_1 \mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \tag{16}
\]
Step II: Margin Disparity Discrepancy (MDD)

- Margin Disparity: \( \epsilon_D^{(\rho)}(f', f) \triangleq \mathbb{E}_{x \sim D_X}[\Phi_\rho(\rho f'(x), h_f(x))]. \)
- 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

\[
\begin{align*}
    d^{(\rho)}_{f, \mathcal{F}}(P, Q) &\triangleq \sup_{f' \in \mathcal{F}} \left( \epsilon^{(\rho)}_Q(f', f) - \epsilon^{(\rho)}_P(f', f) \right), \\
    d^{(\rho)}_{f, \mathcal{F}}(\hat{P}, \hat{Q}) &\triangleq \sup_{f' \in \mathcal{F}} \left( \epsilon^{(\rho)}_{\hat{Q}}(f', f) - \epsilon^{(\rho)}_{\hat{P}}(f', f) \right). 
\end{align*}
\]

(17)

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

---

Margin Theory for Transfer Learning

Theorem (Generalization Bound with Rademacher Complexity)

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

\[
\epsilon_Q(f) \leq \epsilon_{P}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\hat{P}, \hat{Q}) + \epsilon_{\text{ideal}} \\
+ \frac{2k^2}{\rho} \mathcal{R}_{n,P}^{\mathcal{F}}(\Pi_1^\mathcal{F}) + \frac{k}{\rho} \mathcal{R}_{n,P}^{\mathcal{H}\mathcal{F}}(\Pi_1^\mathcal{H}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\
+ \frac{k}{\rho} \mathcal{R}_{m,Q}^{\mathcal{H}\mathcal{F}}(\Pi_1^\mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.
\]

(18)

An expected observation is that the generalization risk is controlled by $\rho$. 

Margin Theory for Transfer Learning

Theorem (Generalization Bound with Covering Numbers)

Let $\mathcal{F} \subseteq \mathbb{R}^{X \times Y}$ be a hypothesis set with label set $Y = \{1, \cdots, k\}$ and $\mathcal{H} \subseteq Y^X$ be the corresponding $Y$-valued labeling function class. Suppose $\Pi_1 \mathcal{F}$ is bounded in $L_2$ by $L$. Fix $\rho > 0$. For all $\delta > 0$, with probability $1 - 3\delta$ the following inequality holds for all hypothesis $f \in \mathcal{F}$:

$$
\epsilon_Q(f) \leq \epsilon_P^{(\rho)}(f) + d_{f, \mathcal{F}}^{(\rho)}(\hat{P}, \hat{Q}) + \epsilon_{\text{ideal}} + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}}
$$

$$
+ \sqrt{\log \frac{2}{\delta}} + \frac{16k^2\sqrt{k}}{\rho} \inf_{\epsilon \geq 0} \left\{ \epsilon + 3\left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right) \right\}
$$

$$(\int_{\epsilon}^{L} \sqrt{\log \mathcal{N}_2(\tau, \Pi_1 \mathcal{F})} d\tau + L \int_{\epsilon/L}^{1} \sqrt{\log \mathcal{N}_2(\tau, \Pi_1 \mathcal{H})} d\tau) .
$$

(19)

The margin bound for OOD has same order with the margin bound for IID.
Margin Theory Implied Algorithm (MDD)\textsuperscript{8}

Minimax domain adaptation implied directly through the margin theory

\[
\min_{f,\psi} \epsilon^{(\rho)}_{\psi(\hat{P})}(f) + \left( \epsilon^{(\rho)}_{\psi(\hat{Q})}(f^*, f) - \epsilon^{(\rho)}_{\psi(\hat{P})}(f^*, f) \right)
\]

\[
f^* = \max_{f'} \left( \epsilon^{(\rho)}_{\psi(\hat{Q})}(f', f) - \epsilon^{(\rho)}_{\psi(\hat{P})}(f', f) \right)
\]

\text{Theory}

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

\text{Algorithm}

\textbf{Bridge the Gap}

\textsuperscript{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 $Q$.

$$\Delta H \Delta H$$ - Divergence

$$\text{Disparity Discrepancy}$$

Harder transfer

Easier transfer
Localization for Discrepancies

Supremum over all pairs

Supremum over single space

ℋΔℋ-Divergence

Disparity Discrepancy

Supremum over localized space

Localized ℋΔℋ-Divergence

Localized Disparity Discrepancy
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

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

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

**Definition (Localized Disparity Discrepancy (LDD))**
For $h \in \mathcal{H}$, the **localized disparity discrepancy** from $P$ to $Q$ is defined as

$$d_{h, \mathcal{H}_r}(P, Q) = \sup_{h' \in \mathcal{H}_r} \left( \mathbb{E}_Q L(h', h) - \mathbb{E}_P L(h', h) \right).$$  
(23)
Recall the generalization bound induced by previous discrepancies:

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

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

\[ \text{err}_Q(\hat{h}) \leq \text{err}_{\hat{P}}(\hat{h}) + d_{\mathcal{H}_r\Delta \mathcal{H}_r}(\hat{P}, \hat{Q}) + \lambda + O\left(\frac{d \log n}{n} + \frac{d \log m}{m}\right) \]
\[ + O\left(\sqrt{\frac{2rd \log n}{n}} + \sqrt{\frac{(d_{\mathcal{H}_r\Delta \mathcal{H}_r}(\hat{P}, \hat{Q}) + 2r)d \log m}{m}}\right). \]  

(24)

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

Localization Theory for Transfer Learning

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

$$
\min_{\tilde{h} \in \mathcal{H}} \{\text{err} \tilde{\rho} (\tilde{h}) + d_{\tilde{h},\mathcal{H}_r} (\hat{P}, \hat{Q})\} \leq \min_{\hat{h} \in \mathcal{H}} \text{err} \hat{\rho} (\hat{h}) + d_{\mathcal{H}_r \Delta \mathcal{H}_r} (\hat{P}, \hat{Q})
$$

(25)

**Theorem (Generalization bound with localized disparity discrepancy)**

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

$$
\text{err} Q (\hat{h}) \leq \text{err} \tilde{\rho} (\tilde{h}) + d_{\tilde{h},\mathcal{H}_r} (\hat{P}, \hat{Q}) + \lambda + O \left( \frac{d \log n}{n} + \frac{d \log m}{m} \right)
$$

$$
+ O \left( \sqrt{\frac{(\text{err} \tilde{\rho} (\tilde{h}) + r) d \log n}{n}} + \sqrt{\frac{(\text{err} \tilde{\rho} (\tilde{h}) + d_{\tilde{h},\mathcal{H}_r} (\hat{P}, \hat{Q}) + r) d \log m}{m}} \right).
$$

(26)

---

Outline

1 Transfer Learning

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

3 Open Library
   - Transfer-Learning-Library
# Design Patterns

<table>
<thead>
<tr>
<th>Reproducible</th>
<th>Stable</th>
<th>Extendible</th>
<th>Ease of Use</th>
<th>TorchVision</th>
<th>Documentation</th>
</tr>
</thead>
</table>

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

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

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

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

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