Transfer Learning: Theories and Algorithms

Mingsheng Long

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

mingsheng@tsinghua.edu.cn
http://ise.thss.tsinghua.edu.cn/~mlong

Outline

1. Transfer Learning
2. $\mathcal{H}\Delta\mathcal{H}$-Divergence
   - DAN: Deep Adaptation Network
   - DANN: Domain Adversarial Neural Network
   - MCD: Maximum Classifier Discrepancy
3. Margin Disparity Discrepancy
   - MDD: Margin Disparity Discrepancy
4. Transfer Model Selection
   - DEV: Deep Embedded Validation
5. Evaluation and Implementation

Supervised Learning

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

fish

bird

mammal

tree

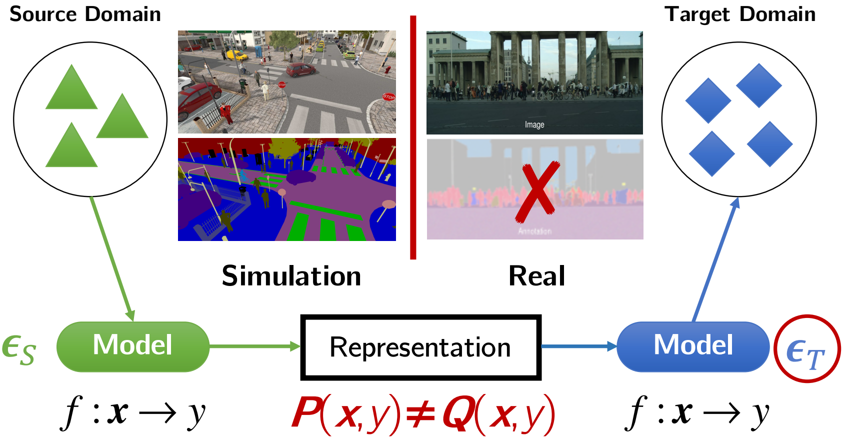
flower

......

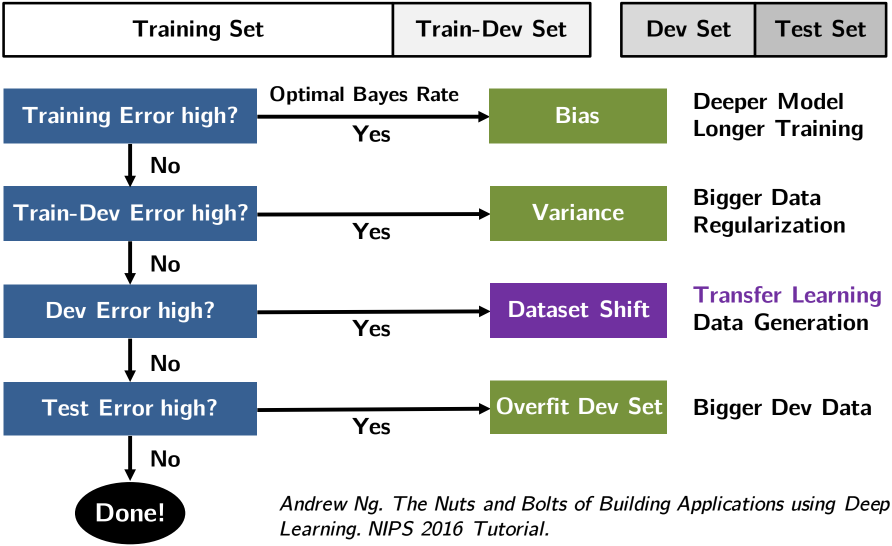
Error Bound: $\epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{complexity}}{n}}$

Transfer Learning

- Machine learning across domains of different distributions $P \neq Q$
  - Independent and Differently Distributed (**IDD**)
- How to effectively bound the generalization error on target domain?

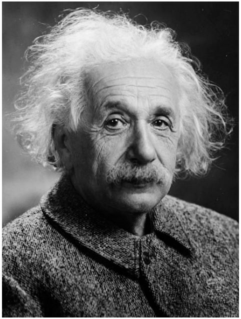


Bias-Variance-Shift Tradeoff



Andrew Ng. The Nuts and Bolts of Building Applications using Deep Learning. NIPS 2016 Tutorial.

Bridging Theory and Algorithm



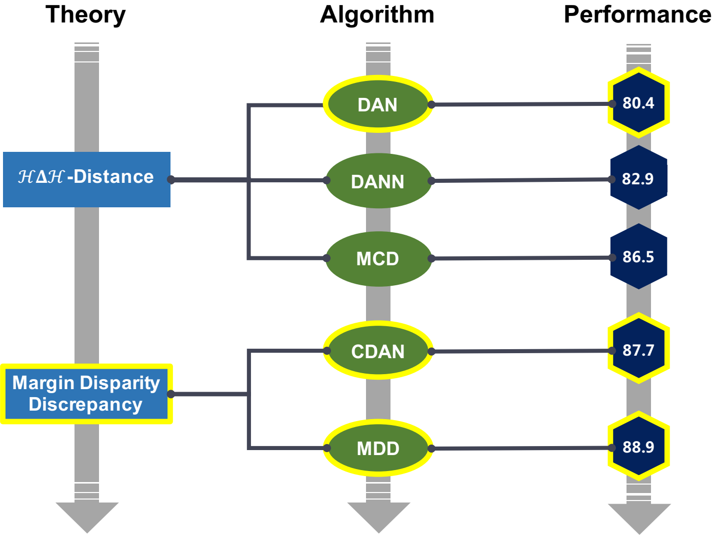
Everything should be made as simple as possible, but no simpler.

—Albert Einstein

There is nothing more practical than a good theory.

—Vladimir Vapnik

Bridging Theory and Algorithm



Outline

1. Transfer Learning

2. $\mathcal{H}\Delta\mathcal{H}$-Divergence
   - DAN: Deep Adaptation Network
   - DANN: Domain Adversarial Neural Network
   - MCD: Maximum Classifier Discrepancy

3. Margin Disparity Discrepancy
   - MDD: Margin Disparity Discrepancy

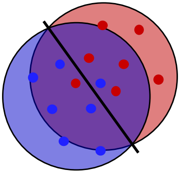
4. Transfer Model Selection
   - DEV: Deep Embedded Validation

5. Evaluation and Implementation

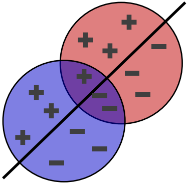
Notations and Assumptions

- Source risk: $\epsilon_P(h) = \mathbb{E}_{(\mathbf{x},y)\sim P}[h(\mathbf{x}) \neq y]$, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim P^n$
- Target risk: $\epsilon_Q(h) = \mathbb{E}_{(\mathbf{x},y)\sim Q}[h(\mathbf{x}) \neq y]$, $\{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim Q^m$
- Source disparity: $\epsilon_P(h_1, h_2) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$
- Target disparity: $\epsilon_Q(h_1, h_2) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$

- Ideal joint hypothesis: $h^* = \arg\min_h \epsilon_P(h) + \epsilon_Q(h)$
- Assumption: ideal hypothesis has small risk $\epsilon_{ideal} = \epsilon_P(h^*) + \epsilon_Q(h^*)$



Distribution discrepancy



Ideal hypothesis with small error

Relating the Target Risk to the Source Risk

Theorem

Assuming small ϵ_{ideal}, the bound of the target risk $\epsilon_Q(h)$ of hypothesis $h \in \mathcal{H}$ is given by the source risk $\epsilon_P(h)$ plus 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.

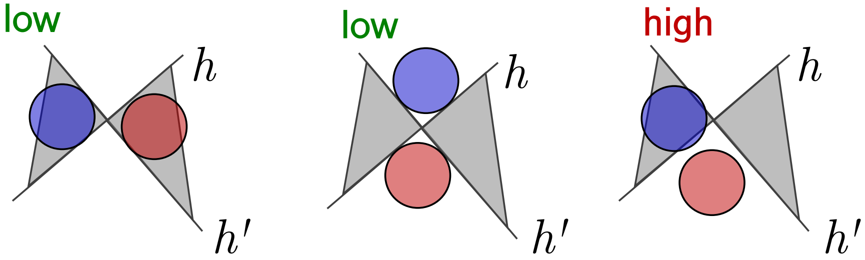
Simply by using the triangle inequalities, we have

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

□

How to Bound the Disparity Difference?

- We can illustrate the disparity difference $\left|\epsilon_P\left(h, h^*\right) - \epsilon_Q\left(h, h^*\right)\right|$ as



- $\mathcal{H}\Delta\mathcal{H}$-Divergence[1]: $d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) \triangleq \sup\limits_{h,h' \in \mathcal{H}} \left|\epsilon_P\left(h, h'\right) - \epsilon_Q\left(h, h'\right)\right|$
- Hypothesis-independent discrepancy—depending on hypothesis space.

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

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

Theorem (Generalization Bound)

Denote by d the VC-dimension of hypothesis space $\mathcal{H}$. For any hypothesis $h \in \mathcal{H}$,

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

- $\epsilon_P(h)$ depicts the performance of h on source domain.
- $d_{\mathcal{H}\Delta\mathcal{H}}$ bounds the generalization gap caused by domain shift.
- ϵ_{ideal} quantifies the inverse of "*adaptability*" between domains.
- The order of the complexity term is $O(\sqrt{d \log n/n} + \sqrt{d \log m/m})$.

Approximating $\mathcal{H}\Delta\mathcal{H}$-Divergence by Statistical Distance

For binary hypothesis h, the $\mathcal{H}\Delta\mathcal{H}$-Divergence can be bounded by

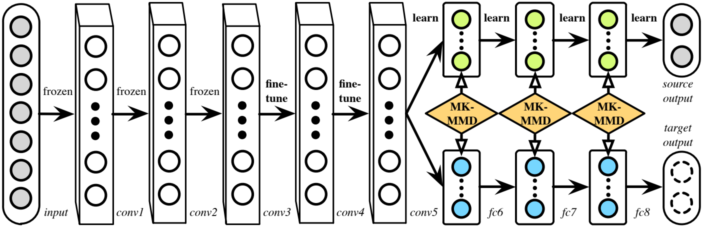
$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) &\triangleq \sup_{h,h'\in\mathcal{H}} \left|\epsilon_P\left(h,h'\right)-\epsilon_Q\left(h,h'\right)\right| \\ &= \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_P\left[\left|h\left(\mathbf{x}\right)-h'\left(\mathbf{x}\right)\right|\neq 0\right]-\mathbb{E}_Q\left[\left|h\left(\mathbf{x}\right)-h'\left(\mathbf{x}\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

The last term takes the form of Integral Probability Metric (IPM):

$$d_{\mathcal{F}}(P,Q)=\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{\mathbf{x}\sim P}f(\mathbf{x})-\mathbb{E}_{\mathbf{x}\sim Q}f(\mathbf{x})\right| \tag{5}$$

Assuming $\mathcal{F}$ can be approximated by kernel functions in RKHS, $d_{\mathcal{F}}(P,Q)$ turns into Maximum Mean Discrepancy (MMD) (a statistical distance)

DAN: Deep Adaptation Network[2]



Distribution matching: yield the upper-bound by multiple kernel learning

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

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

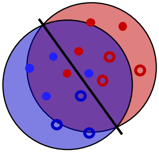
[2] *Long et al. Learning Transferable Features with Deep Adaptation Networks. ICML 2015.*

Approximating $\mathcal{H}\Delta\mathcal{H}$-Divergence by Domain Discriminator

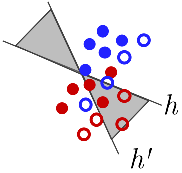
For binary hypothesis h, the $\mathcal{H}\Delta\mathcal{H}$-Divergence can be bounded by

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) &\triangleq \sup_{h,h' \in \mathcal{H}} \left|\epsilon_P\left(h, h'\right) - \epsilon_Q\left(h, h'\right)\right| \\ &= \sup_{\delta \in \mathcal{H}\Delta\mathcal{H}} \left|\mathbb{E}_P\left[\delta(\mathbf{x}) \neq 0\right] - \mathbb{E}_Q\left[\delta\left(\mathbf{x}\right) \neq 0\right]\right| \\ &\leqslant \sup_{D \in \mathcal{H}_D} \left|\mathbb{E}_P\left[D(\mathbf{x}) = 1\right] + \mathbb{E}_Q\left[D\left(\mathbf{x}\right) = 0\right]\right| \end{aligned} \tag{8}$$

This upper-bound can be yielded by training a domain discriminator $D(\mathbf{x})$

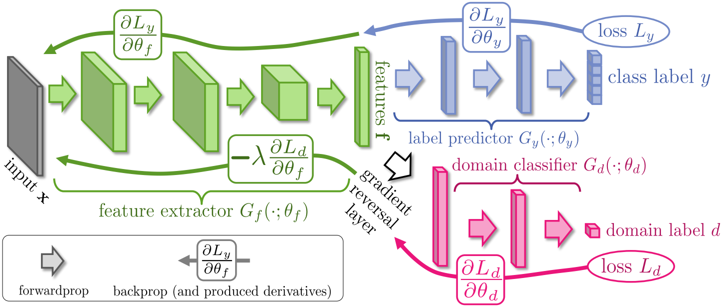


Distribution discrepancy



Hypothesis-based distribution discrepancy

DANN: Domain Adversarial Neural Network[3]



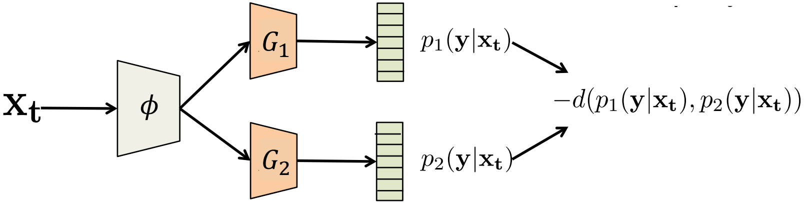
Adversarial adaptation: learning features indistinguishable across domains

$$E(\theta_f, \theta_y, \theta_d) = \sum_{\mathbf{x}_i \sim \widehat{P}} L_y(G_y(G_f(\mathbf{x}_i)), y_i) - \lambda \sum_{\mathbf{x}_i \sim \widehat{P} \cup \widehat{Q}} L_d(G_d(G_f(\mathbf{x}_i)), d_i) \quad (9)$$

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \theta_d) \quad (\hat{\theta}_d) = \arg\max_{\theta_d} E(\theta_f, \theta_y, \theta_d) \quad (10)$$

[3] *Ganin et al. Domain Adversarial Training of Neural Networks. JMLR 2016.*

Approximating $\mathcal{H}\Delta\mathcal{H}$-Divergence by Classifier Consistency[4]



- Use two classifiers G_1, G_2 to approximate $\sup_{h,h' \in \mathcal{H}} \left|\epsilon_P(h,h') - \epsilon_Q(h,h')\right|$
- Assume $G_1 = h$ and $G_2 = h'$ should agree on source domain.
- Use L_1-loss of two classifiers' outputs to approximate disagreement:

$$\min_{\phi}\{\min_{G_1,G_2} \mathbb{E}_{\widehat{P}}[L(G_1(\mathbf{x}), y) + L(G_2(\mathbf{x}), y)] + \max_{G_1,G_2} \mathbb{E}_{\widehat{Q}}|G_1(\mathbf{x}) - G_2(\mathbf{x})|\} \quad (11)$$

[4] *Saito, et al. Maximum classifier discrepancy for unsupervised domain adaptation. CVPR 2018.*

Outline

Towards Informative Margin Theory[5]

- Towards a rigorous multiclass domain adaptation theory.
  - All existing theories are only applicable to binary classification.
- Towards an informative margin theory.
  - Explore the idea of margin in measuring domain discrepancy.
- Towards a certain function class in the theoretical bound.
  - Eliminate approximation assumptions in all existing methods.
- Towards bridging the existing gap between theories and algorithms.

[5] *Zhang et al. Bridging Theory and Algorithm for Domain Adaptation. ICML 2019.*

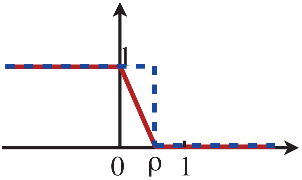
Notations

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

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

- Margin Loss:

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



MDD: Margin Disparity Discrepancy

- Margin risk: $\epsilon_D^{(\rho)}(f) = \mathbb{E}_{(x,y)\sim D}[\Phi_\rho(\rho_f(x,y))]$
- Margin disparity: $\epsilon_D^{(\rho)}(f', f) \triangleq \mathbb{E}_{x\sim D_X}[\Phi_\rho(\rho_{f'}(x, h_f(x)))]$

Definition (Margin Disparity Discrepancy, MDD)

With above definitions, we define Margin Disparity Discrepancy (MDD) and its empirical version by

$$
\begin{aligned}
d_{f,\mathcal{F}}^{(\rho)}(P,Q) &\triangleq \sup_{f'\in\mathcal{F}} \left(\epsilon_Q^{(\rho)}(f',f) - \epsilon_P^{(\rho)}(f',f)\right), \\
d_{f,\mathcal{F}}^{(\rho)}(\widehat{P},\widehat{Q}) &\triangleq \sup_{f'\in\mathcal{F}} \left(\epsilon_{\widehat{Q}}^{(\rho)}(f',f) - \epsilon_{\widehat{P}}^{(\rho)}(f',f)\right).
\end{aligned} \tag{12}
$$

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

Bounding the Target Risk by MDD

Theorem

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. For every scoring function $f \in \mathcal{F}$,

$$\epsilon_Q(f) \leq \epsilon_P^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(P, Q) + \epsilon_{ideal}^{(\rho)}, \tag{13}$$

where $\epsilon_{ideal}^{(\rho)}$ is the margin error of ideal joint hypothesis f^*:

$$\epsilon_{ideal}^{(\rho)} = \min_{f^* \in \mathcal{F}} \{\epsilon_P^{(\rho)}(f^*) + \epsilon_Q^{(\rho)}(f^*)\}. \tag{14}$$

- Main proof difficulties: margin loss does not satisfy triangle inequality.
  - Solution: One-sided triangle inequality for the margin loss.
- **A new tool for analyzing transfer learning with margin theory.**

Definitions

Definition (Function Class $\Pi_1\mathcal{F}$)

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

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

Definition (Function Class $\Pi_{\mathcal{H}}\mathcal{F}$)

Given a class of scoring functions $\mathcal{F}$ and a class of the induced labeling functions $\mathcal{H}$, we define $\Pi_{\mathcal{H}}\mathcal{F}$ as

$$\Pi_{\mathcal{H}}\mathcal{F} \triangleq \{x \mapsto f(x, h(x)) | h \in \mathcal{H}, f \in \mathcal{F}\}. \tag{16}$$

By applying the margin error over each entry in $\Pi_{\mathcal{H}}\mathcal{F}$, we obtain the **"scoring" version** of $\mathcal{H}\Delta\mathcal{H}$ (symmetric difference hypothesis space)

Definitions

Definition (Rademacher Complexity)

The empirical Rademacher complexity of function class $\mathcal{G}$ with respect to the sample $\widehat{D}$ is defined as

$$\widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{G}) = \mathbb{E}_{\sigma} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i). \tag{17}$$

where σ_i's are independent uniform random variables taking values in $\{-1, +1\}$. The Rademacher complexity is

$$\mathfrak{R}_{n,D}(\mathcal{G}) = \mathbb{E}_{\widehat{D} \sim D^n} \widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{G}). \tag{18}$$

Definition (Covering Number)

(Informal) A *covering number* $\mathcal{N}_2(\tau, \mathcal{G})$ is the minimal number of $\mathcal{L}_2$ balls of radius $\tau > 0$ needed to cover a class $\mathcal{G}$ of bounded functions $g : \mathcal{X} \to \mathbb{R}$

Generalization Bound with Rademacher Complexity

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 $f \in \mathcal{F}$:

$$
\begin{aligned}
\epsilon_Q(f) \leq & \epsilon_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} \\
& + \frac{2k^2}{\rho}\mathfrak{R}_{n,P}(\Pi_1\mathcal{F}) + \frac{k}{\rho}\mathfrak{R}_{n,P}(\Pi_{\mathcal{H}}\mathcal{F}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\
& + \frac{k}{\rho}\mathfrak{R}_{m,Q}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.
\end{aligned}
\tag{19}
$$

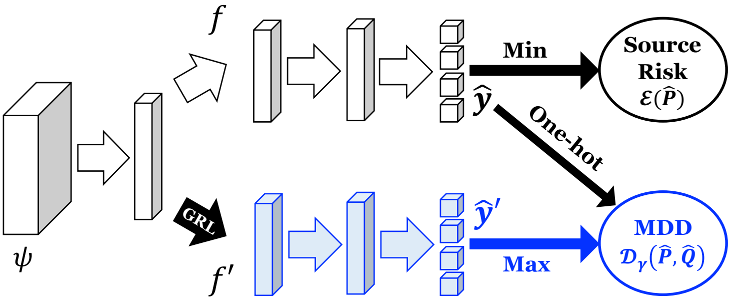
Generalization Bound with Covering Numbers

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 $f \in \mathcal{F}$:

$$\begin{aligned}\epsilon_Q(f) \leq & \epsilon_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\ & + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \frac{16k^2\sqrt{k}}{\rho}\inf_{\epsilon\geq 0}\Big\{\epsilon + 3\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big) \\ & \Big(\int_{\epsilon}^{L}\sqrt{\log\mathcal{N}_2(\tau,\Pi_1\mathcal{F})}\mathrm{d}\tau + L\int_{\epsilon/L}^{1}\sqrt{\log\mathcal{N}_2(\tau,\Pi_1\mathcal{H})}\mathrm{d}\tau\Big)\Big\}.\end{aligned} \tag{20}$$

MDD: Margin Disparity Discrepancy



Minimax game: Adversarial learning induced by informative margin theory

$$\begin{aligned} &\min_{f,\psi} \ \epsilon^{(\rho)}_{\psi(\widehat{P})}(f) + (\epsilon^{(\rho)}_{\psi(\widehat{Q})}(f^*, f) - \epsilon^{(\rho)}_{\psi(\widehat{P})}(f^*, f)), \\ &f^* = \max_{f'} (\epsilon^{(\rho)}_{\psi(\widehat{Q})}(f', f) - \epsilon^{(\rho)}_{\psi(\widehat{P})}(f', f)). \end{aligned} \tag{21}$$

Outline

Model Selection in Domain Adaptation

- Supervised Learning

$(x_1, y_1) \sim p$ — Training

$(x_2, y_2) \sim p$ — Validation

$(x_3, y_3) \sim p$ — Test

- Semi-Supervised Learning (SSL)?
- Unsupervised Domain Adaptation (UDA)?

Source Domain

$(x_1, y_1) \sim p$

(x_1, y_1)

Training

Validation

x_2

Target Domain

$(x_2, y_2) \sim q$

(x_2, y_2)

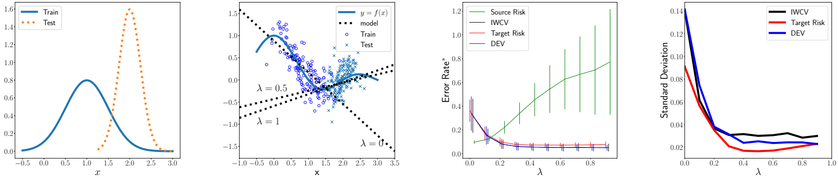
Test

IWCV: Importance-Weighted Cross-Validation[6]

- Covariate shift assumption: $P(y|\mathbf{x}) = Q(y|\mathbf{x})$
- Model selection by estimating Target Risk $\epsilon_Q(h) = \mathbb{E}_Q[h(\mathbf{x}) \neq y]$
- Importance-Weighted Cross-Validation (IWCV)

$$\mathbb{E}_P w(\mathbf{x}) \cdot [h(\mathbf{x}) \neq y] = \mathbb{E}_P \frac{Q(\mathbf{x})}{P(\mathbf{x})} \cdot [h(\mathbf{x}) \neq y] = \mathbb{E}_Q [h(\mathbf{x}) \neq y] = \epsilon_Q(h)$$

  - The estimation is unbiased but the variance is unbounded
  - Density ratio is not accessible due to unknownness of P and Q



[6] *Covariate shift adaptation by importance weighted cross validation, JMLR'2007*

DEV: Deep Embedded Validation[10]

- Variance of IWCV (bounded by Rényi divergence)[7]:

$$\mathrm{Var}_{\mathbf{x}\sim P}[w(\mathbf{x}) \cdot [h(\mathbf{x}) \neq y]] \leq d_{\alpha+1}(Q\|P)\epsilon_Q(h)^{1-\frac{1}{\alpha}} - \epsilon_Q(h)^2$$

- Density ratio $w(\mathbf{x}) = \frac{Q(\mathbf{x})}{P(\mathbf{x})}$ is estimated by discriminative learning[8]
- Feature adaptation reduces distribution discrepancy $d_{\alpha+1}(Q\|P)$[9]
- Control variate explicitly reduces the variance of $\mathbb{E}_P w(\mathbf{x}) \cdot [h(\mathbf{x}) \neq y]$
  - $\mathbb{E}[z] = \zeta, \mathbb{E}[t] = \tau$
  - $z^\star = z + \eta(t - \tau)$
  - $\mathbb{E}[z^\star] = \mathbb{E}[z] + \eta\mathbb{E}[t - \tau] = \zeta + \eta(\mathbb{E}[t] - \mathbb{E}[\tau]) = \zeta.$
  - $\mathrm{Var}[z^\star] = \mathrm{Var}[z + \eta(t - \tau)] = \eta^2\mathrm{Var}[t] + 2\eta\mathrm{Cov}(z, t) + \mathrm{Var}[z]$
  - min $\mathrm{Var}[z^\star] = (1 - \rho^2_{z,t})\mathrm{Var}[z]$, when $\hat{\eta} = -\frac{\mathrm{Cov}(z,t)}{\mathrm{Var}[t]}$

---

[7] *Learning Bounds for Importance Weighting, NeurIPS 2010*

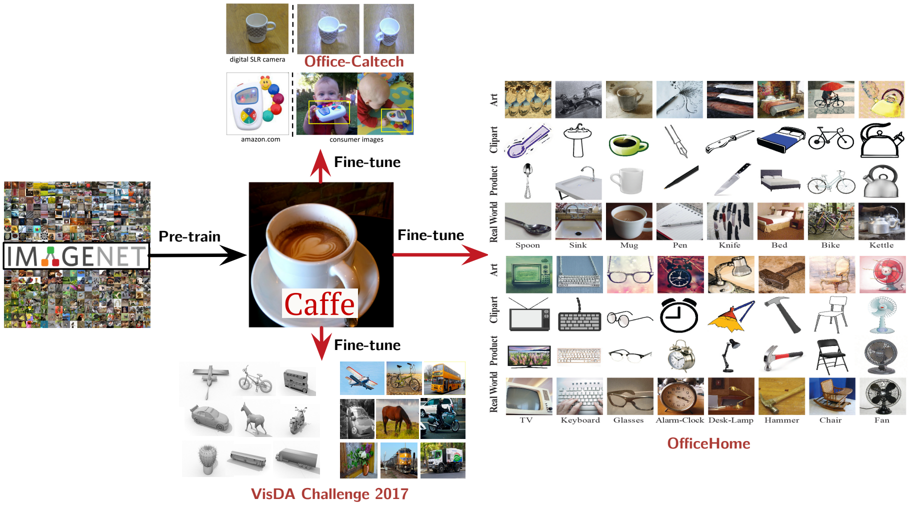
[8] *Discriminative learning for differing training and test distributions, ICML 2007*

[9] *Conditional Adversarial Domain Adaptation, NeurIPS 2018*

[10] *You et al. Towards Accurate Model Selection in Deep Unsupervised Domain Adaptation. ICML 2019.*

Outline

Datasets



digital SLR camera
Office-Caltech
amazon.com
consumer images
Fine-tune
IMAGENET
Pre-train
Caffe
Fine-tune
Fine-tune
Art
Clipart
Product
Real World
Spoon
Sink
Mug
Pen
Knife
Bed
Bike
Kettle
Art
Clipart
Product
Real World
TV
Keyboard
Glasses
Alarm-Clock
Desk-Lamp
Hammer
Chair
Fan
OfficeHome
VisDA Challenge 2017

Results

Table: Accuracy (%) on Office-31 for unsupervised domain adaptation

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50	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	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	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
CDAN	93.0±0.2	98.4±0.2	**100.0**±.0	89.2±0.3	70.2±0.4	69.4±0.4	86.7
CDAN+E	93.1±0.1	**98.6**±0.1	**100.0**±.0	93.4±0.2	71.0±0.3	70.3±0.3	87.7
MDD	**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**

Table: Accuracy (%) on Office-Home for unsupervised domain adaptation

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
ResNet-50	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	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	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
CDAN	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
MDD	**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**

Results

Table: Accuracy (%) on VisDA-2017 (ResNet-50)

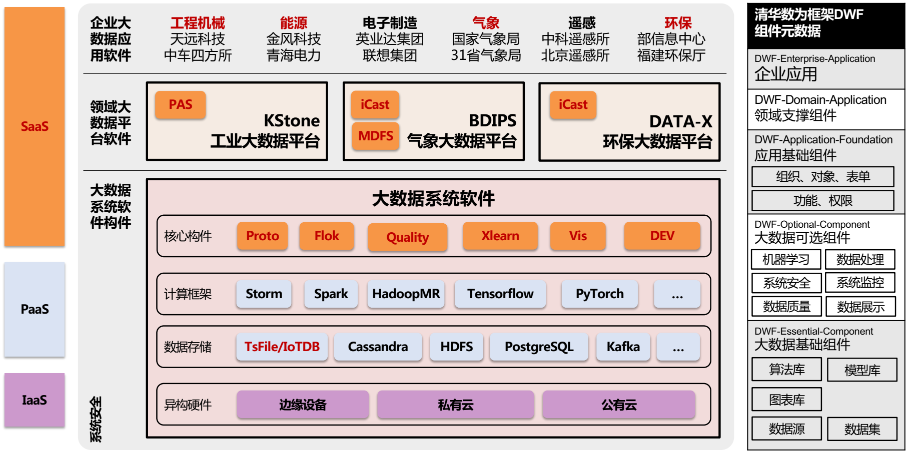
Method	Synthetic → Real
MCD	69.2
GTA	69.5
CDAN	70.0
MDD	**74.6**

Table: Accuracy (%) of **MCD** by different validation methods on VisDA-2017

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	mean
Original	87.00	60.90	83.70	64.00	88.90	79.60	84.70	76.90	88.60	40.30	83.00	25.80	71.90
Source Risk	84.39	54.11	69.15	46.37	80.49	80.45	85.04	65.24	87.22	36.86	78.04	28.91	66.36
IWCV	81.21	60.95	76.00	56.53	82.83	72.06	84.05	68.65	86.85	44.37	69.29	23.81	67.22
DEV (w/o control variate)	84.21	63.95	79.00	59.53	85.83	75.06	87.05	71.65	89.85	47.37	72.29	26.81	70.22
DEV	81.83	53.48	82.95	71.62	89.16	72.03	89.36	75.73	97.02	55.48	71.19	29.17	72.42
Target Risk (Upper Bound)	81.95	53.60	83.07	72.02	89.25	72.15	89.55	75.83	97.10	55.57	71.19	29.27	72.55

Transfer Learning Systems

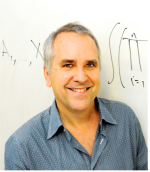
Tsinghua Dataway Big Data Software Stack



National Engineering Lab for Big Data Software

Jiaguang Sun (孙家广)
Director
Tsinghua University

Jianmin Wang (王建民)
Deputy Director
Dean, School of Software
Tsinghua University

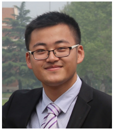


Michael I. Jordan
Academic Committee Chair
UC Berkeley

Machine Learning Group @ NELBDS

Mingsheng Long (龙明盛)

Yuchen Zhang



Yue Cao

Han Zhu

Zhangjie Cao