Transfer Learning via Learning to Transfer

Ying WEI, Yu Zhang, Junzhou Huang, Qiang Yang

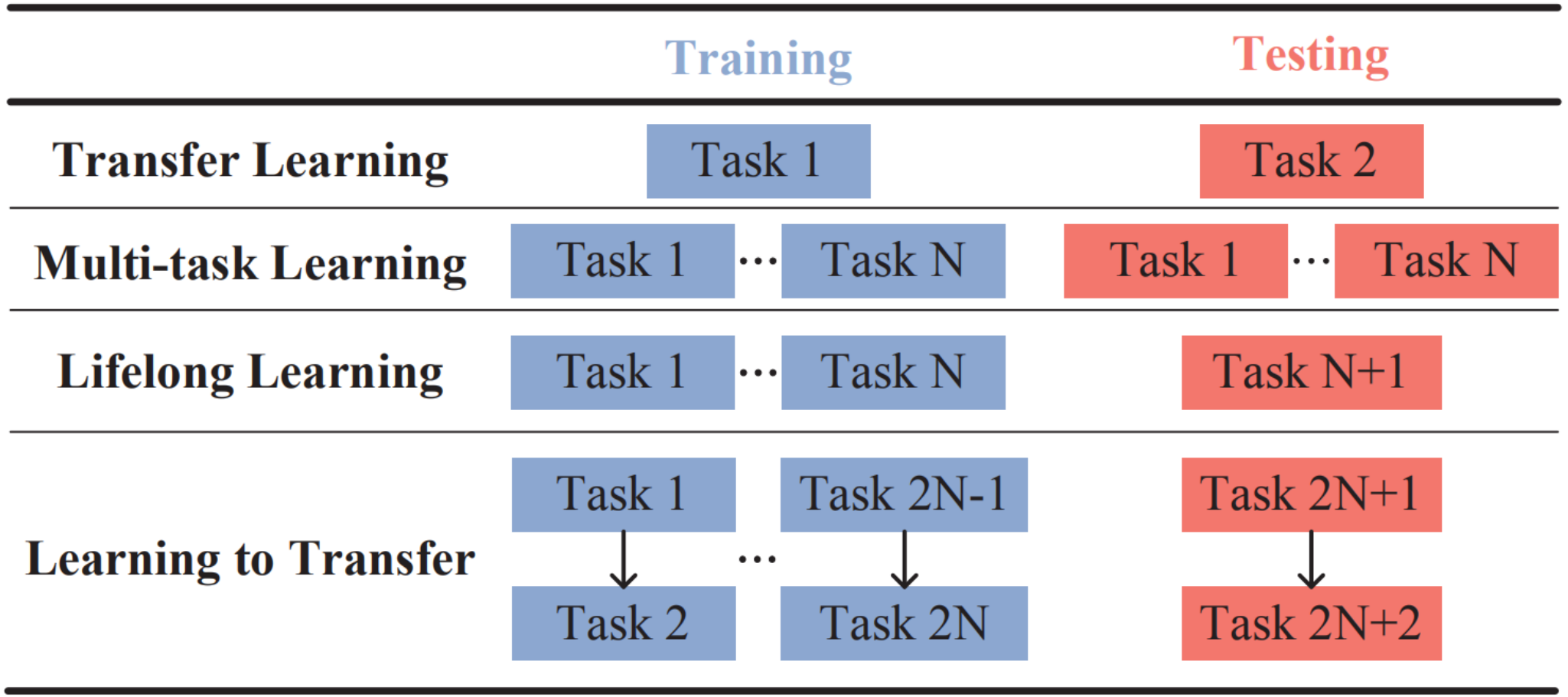
ICML2018

Introduction

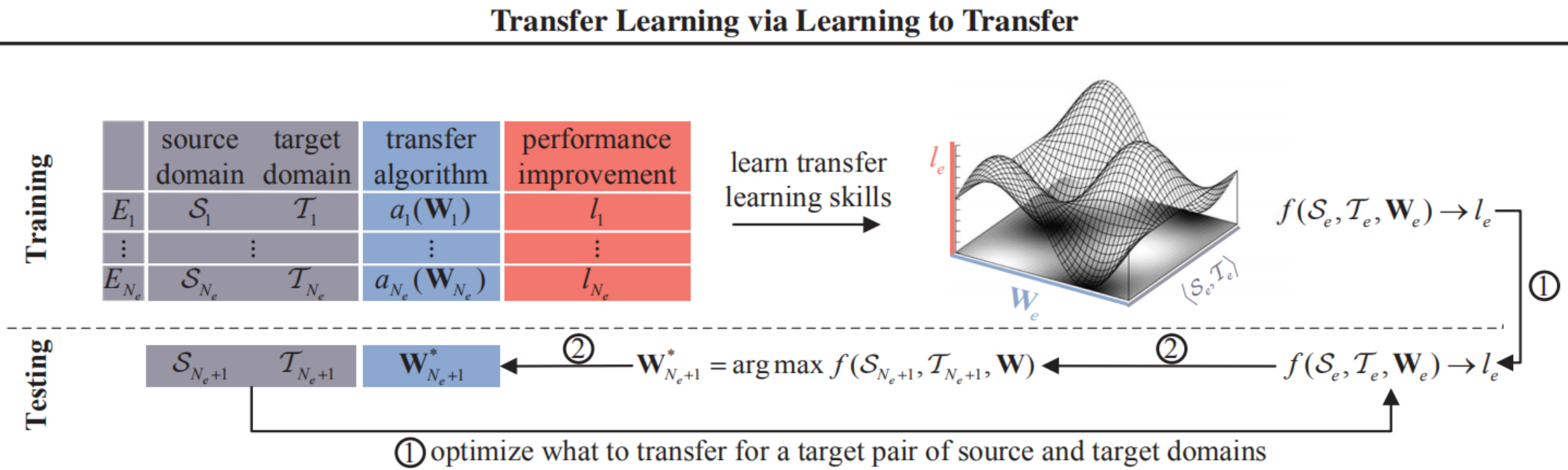
Definition 1 *(Transfer Learning)* Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, *transfer learning* aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.

$$\mathcal{D} = \{\mathcal{X}, P(X)\} \qquad \mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$$

Introduction



Framework



$$\mathcal{S}_e = \{\mathbf{X}_e^s, \mathbf{y}_e^s\} \; \mathcal{T}_e = \{\mathbf{X}_e^t, \mathbf{y}_e^t\} \qquad l_e = p_e^{st}/p_e^t$$

$$\mathbf{X}_e^* \in \mathbb{R}^{n_e^* \times m}$$

What to Transfer

a latent factor matrix feature **W**

Latent feature factor based algorithms aim to learn domain-invariant feature factors across domains.

$$Z=XW$$

Learning from Experiences

$$f(\mathcal{S}_e, \mathcal{T}_e, \mathbf{W}_e) \rightarrow l_e$$

The Difference between a Source and a Target Domain

The Discriminative Ability of a Target Domain

Difference

maximum mean discrepancy (MMD)

$$
\begin{aligned}
&\hat{d}_e^2(\mathbf{X}_e^s\mathbf{W}_e, \mathbf{X}_e^t\mathbf{W}_e) \\
=&\left\| \frac{1}{n_e^s}\sum_{i=1}^{n_e^s}\phi(\mathbf{x}_{ei}^s\mathbf{W}_e) - \frac{1}{n_e^t}\sum_{j=1}^{n_e^t}\phi(\mathbf{x}_{ej}^t\mathbf{W}_e)\right\|_{\mathcal{H}}^2 \\
=&\frac{1}{(n_e^s)^2}\sum_{i,i'=1}^{n_e^s}\mathcal{K}(\mathbf{x}_{ei}^s\mathbf{W}_e, \mathbf{x}_{ei'}^s\mathbf{W}_e) \\
+&\frac{1}{(n_e^t)^2}\sum_{j,j'=1}^{n_e^t}\mathcal{K}(\mathbf{x}_{ej}^t\mathbf{W}_e, \mathbf{x}_{ej'}^t\mathbf{W}_e) \\
-&\frac{2}{n_e^s n_e^t}\sum_{i,j=1}^{n_e^s,n_e^t}\mathcal{K}(\mathbf{x}_{ei}^s\mathbf{W}_e, \mathbf{x}_{ej}^t\mathbf{W}_e),
\end{aligned}
$$

Difference

Variances

$$d_e^2(\mathbf{X}_e^s\mathbf{W}_e, \mathbf{X}_e^t\mathbf{W}_e) = \mathbf{E}_{\mathbf{x}_e^s\mathbf{x}_e^{s\prime}\mathbf{x}_e^t\mathbf{x}_e^{t\prime}} h(\mathbf{x}_e^s, \mathbf{x}_e^{s\prime}, \mathbf{x}_e^t, \mathbf{x}_e^{t\prime})$$

$$h(\mathbf{x}_e^s, \mathbf{x}_e^{s\prime}, \mathbf{x}_e^t, \mathbf{x}_e^{t\prime}) = \mathcal{K}(\mathbf{x}_e^s\mathbf{W}_e, \mathbf{x}_e^{s\prime}\mathbf{W}_e) + \mathcal{K}(\mathbf{x}_e^t\mathbf{W}_e, \mathbf{x}_e^{t\prime}\mathbf{W}_e) - \mathcal{K}(\mathbf{x}_e^s\mathbf{W}_e, \mathbf{x}_e^{t\prime}\mathbf{W}_e) - \mathcal{K}(\mathbf{x}_e^{s\prime}\mathbf{W}_e, \mathbf{x}_e^t\mathbf{W}_e)$$

$$\sigma_e^2(\mathbf{X}_e^s\mathbf{W}_e, \mathbf{X}_e^t\mathbf{W}_e) = \mathbf{E}_{\mathbf{x}_e^s\mathbf{x}_e^{s\prime}\mathbf{x}_e^t\mathbf{x}_e^{t\prime}} [(h(\mathbf{x}_e^s, \mathbf{x}_e^{s\prime}, \mathbf{x}_e^t, \mathbf{x}_e^{t\prime}) - \mathbf{E}_{\mathbf{x}_e^s\mathbf{x}_e^{s\prime}\mathbf{x}_e^t\mathbf{x}_e^{t\prime}} h(\mathbf{x}_e^s, \mathbf{x}_e^{s\prime}, \mathbf{x}_e^t, \mathbf{x}_e^{t\prime}))^2]$$

Discriminative

$$\tau_e = \mathrm{tr}(\mathbf{W}_e^T \mathbf{S}_e^N \mathbf{W}_e) / \mathrm{tr}(\mathbf{W}_e^T \mathbf{S}_e^L \mathbf{W}_e)$$

$$\mathbf{S}_e^L = \sum_{j,j'=1}^{n_e^t} \frac{H_{jj'}}{(n_e^t)^2} (\mathbf{x}_{ej}^t - \mathbf{x}_{ej'}^t)(\mathbf{x}_{ej}^t - \mathbf{x}_{ej'}^t)^T$$

$$H_{jj'} = \begin{cases} \mathcal{K}(\mathbf{x}_{ej}^t, \mathbf{x}_{ej'}^t), & \text{if} \quad \mathbf{x}_{ej}^t \in \mathcal{N}_r(\mathbf{x}_{ej'}^t) \quad \text{and} \quad \mathbf{x}_{ej'}^t \in \mathcal{N}_r(\mathbf{x}_{ej}^t) \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{S}_e^N = \sum_{j,j'=1}^{n_e^t} \frac{\mathcal{K}(\mathbf{x}_{ej}^t, \mathbf{x}_{ej'}^t) - H_{jj'}}{(n_e^t)^2} (\mathbf{x}_{ej}^t - \mathbf{x}_{ej'}^t)(\mathbf{x}_{ej}^t - \mathbf{x}_{ej'}^t)^T$$

Learning from Experiences

$$\frac{1}{f} = \boldsymbol{MMD} + \lambda \boldsymbol{Variances} + \frac{\mu}{\boldsymbol{\tau}_e} + b \approx \frac{1}{l_e}$$

Thus learning the reflection function f is equivalent to optimizing kernel function K.

Learning from Experiences

$$\mathcal{K} = \sum_{k=1}^{N_k} \beta_k \mathcal{K}_k \ (\beta_k \geq 0, \forall k)$$

$$1/f = \boldsymbol{\beta}^T \hat{\mathbf{d}}_e + \lambda \boldsymbol{\beta}^T \hat{\mathbf{Q}}_e \boldsymbol{\beta} + \frac{\mu}{\boldsymbol{\beta}^T \boldsymbol{\tau}_e} + b$$

$$\mathbf{Q}_e = \mathrm{cov}(h) = \begin{bmatrix} \sigma_{e(1,1)} & \cdots & \sigma_{e(1,N_k)} \\ \cdots & \cdots & \cdots \\ \sigma_{e(N_k,1)} & \cdots & \sigma_{e(N_k,N_k)} \end{bmatrix}$$

Learning from Experiences

$$\boldsymbol{\beta}^*, \lambda^*, \mu^*, b^* =$$

$$\arg\min_{\boldsymbol{\beta},\lambda,\mu,b} \sum_{e=1}^{N_e} \mathcal{L}_h \left(\boldsymbol{\beta}^T \hat{\mathbf{d}}_e + \lambda \boldsymbol{\beta}^T \hat{\mathbf{Q}}_e \boldsymbol{\beta} + \frac{\mu}{\boldsymbol{\beta}^T \boldsymbol{\tau}_e} + b, \frac{1}{l_e} \right)$$

$$+ \gamma_1 R(\boldsymbol{\beta}, \lambda, \mu, b),$$

$$\text{s.t.} \quad \beta_k \geq 0, \ \forall k \in \{1, \cdots, N_k\}, \ \lambda \geq 0, \ \mu \geq 0$$

Huber regression loss

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, & \text{for } |y - f(x)| \leq \delta \\ \delta \cdot (|y - f(x)| - \delta/2), & \text{otherwise.} \end{cases}$$

Inferring What to Transfer

$$\mathbf{W}^*_{N_e+1} = \arg\max_{\mathbf{W}} f(\mathcal{S}_{N_e+1}, \mathcal{T}_{N_e+1}, \mathbf{W}; \boldsymbol{\beta}^*, \lambda^*, \mu^*, b^*) - \gamma_2 \|\mathbf{W}\|_F^2$$

$$= \arg\min_{\mathbf{W}} (\boldsymbol{\beta}^*)^T \hat{\mathbf{d}}_{\mathbf{W}} + \lambda^* (\boldsymbol{\beta}^*)^T \hat{\mathbf{Q}}_{\mathbf{W}} \boldsymbol{\beta}^* + \mu^* \frac{1}{(\boldsymbol{\beta}^*)^T \boldsymbol{\tau}_{\mathbf{W}}}$$

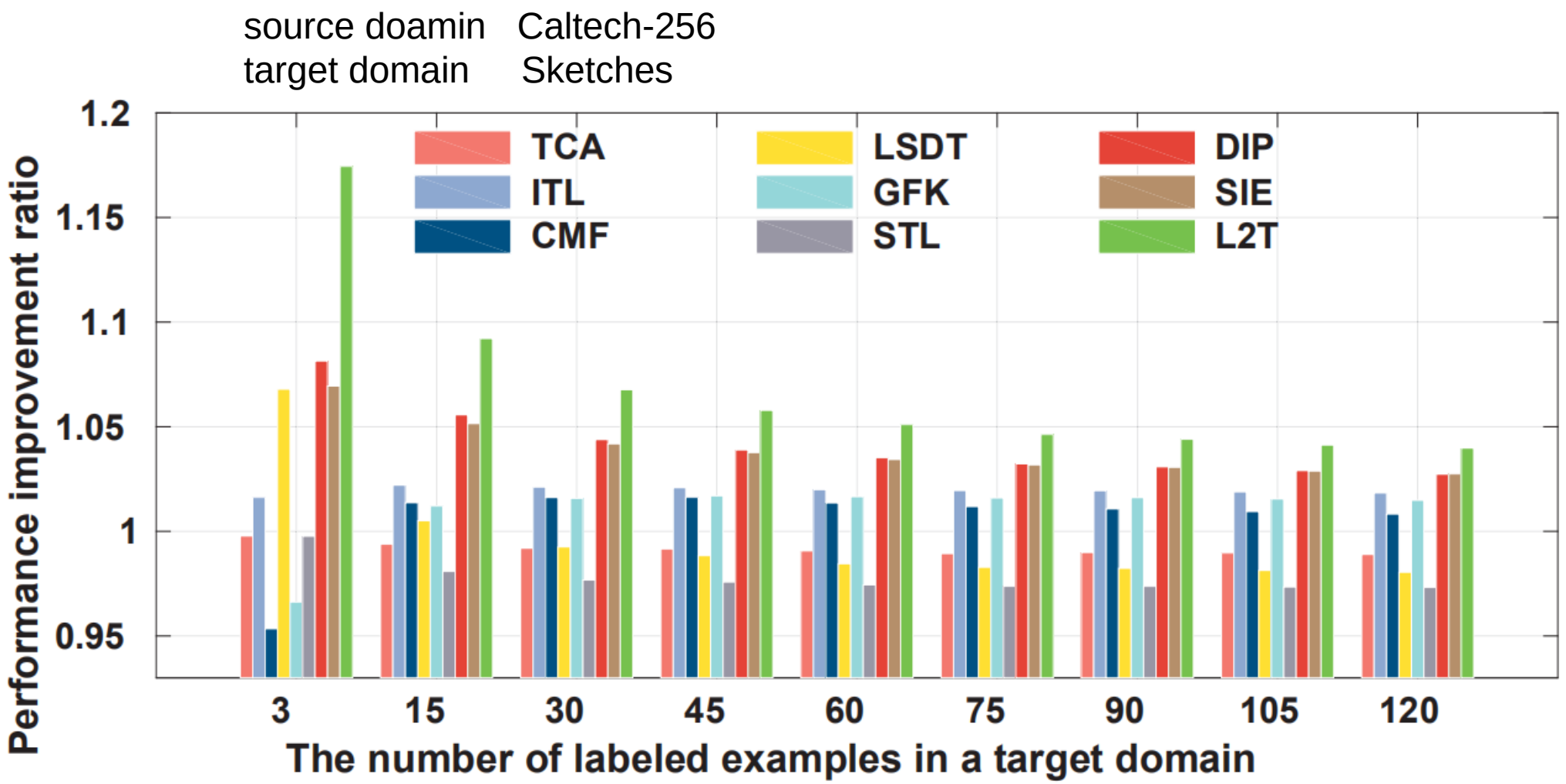
$$+ \gamma_2 \|\mathbf{W}\|_F^2,$$

Theorem 2. *Let* $\delta' = \delta - (N_e)^{\frac{1}{2(r+1)} - \frac{1}{4}}$ *(*$r > 1$ *is required). Then for any sample* $\mathbf{S}$ *of size* N_e *drawn according to an algebraic* β*-mixing stationary distribution, and* $\delta \geq 0$ *such that* $\delta' \geq 0$*, the following generalization bound holds with probability at least* $1 - \delta$*:*

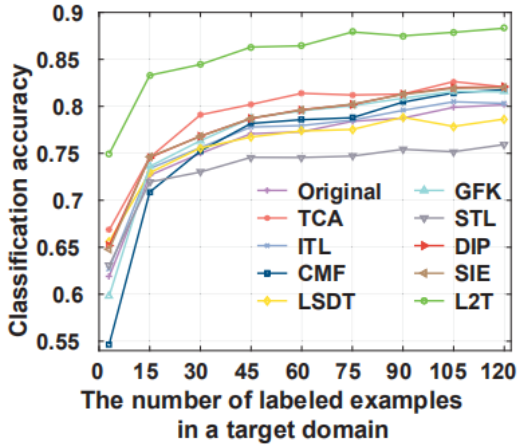
$$\left|R(\mathbf{L}(\mathbf{S})) - R_{N_e}(\mathbf{L}(\mathbf{S}))\right| < \mathcal{O}\left((N_e)^{\frac{1}{2(r+1)} - \frac{1}{4}} \sqrt{\log(\frac{1}{\delta'})}\right),$$

where $R(\mathbf{L}(\mathbf{S}))$ *and* $R_{N_e}(\mathbf{L}(\mathbf{S}))$ *denote the expected risk and the empirical risk of L2T over meta-samples, respectively. A larger mixing parameter* r*, indicating more independence, would lead to a tighter bound.*

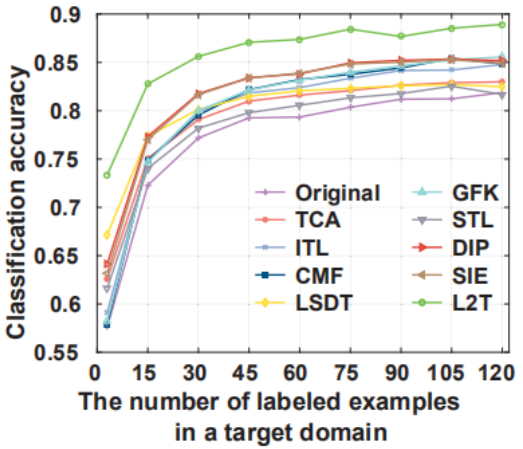
Experiments



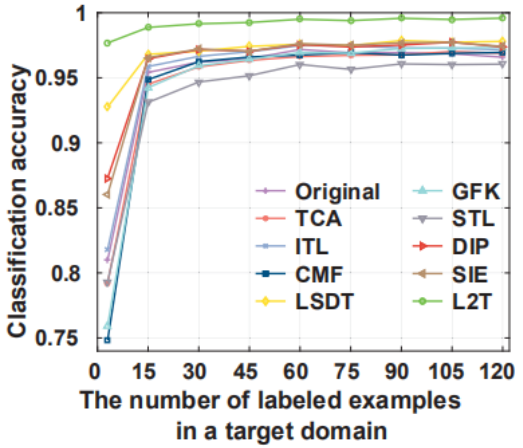
Experiments



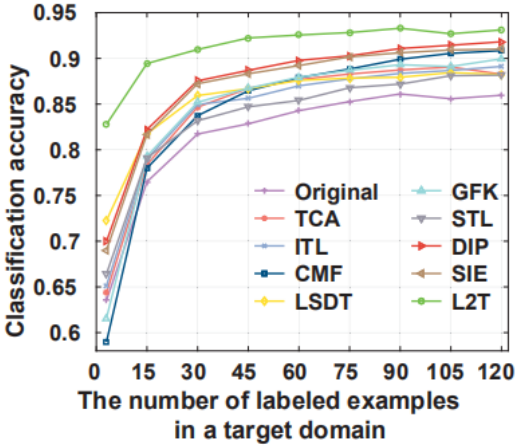
(a) galaxy / harpsichord / saturn → kangaroo / standing-bird / sun



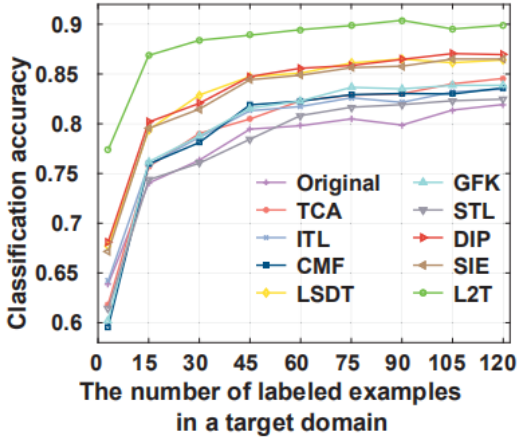
(b) bat / mountain-bike / saddle → bush / person / walkie-talkie



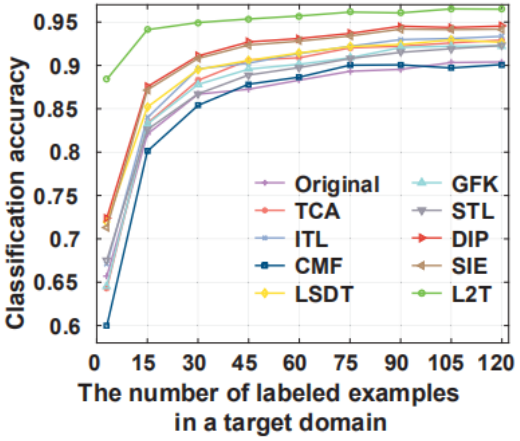
(c) microwave / spider / watch → spoon / trumpet / wheel



(d) bridge / harp / traffic-light → door-handle / hand / present

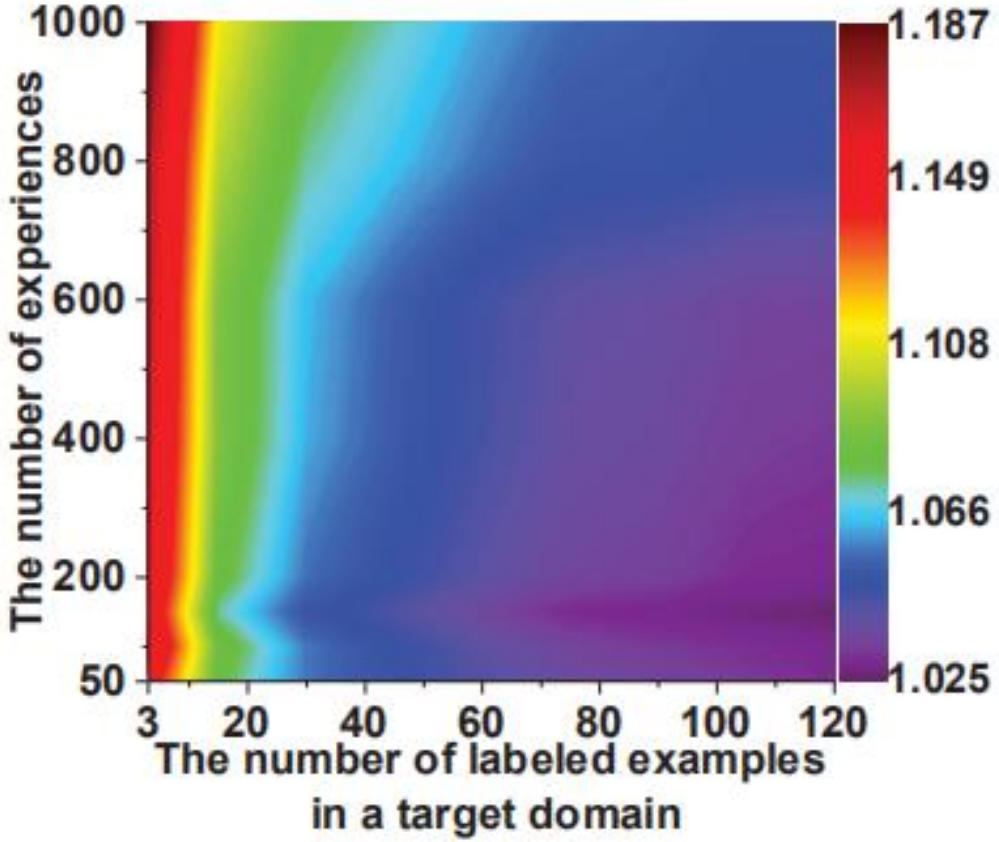


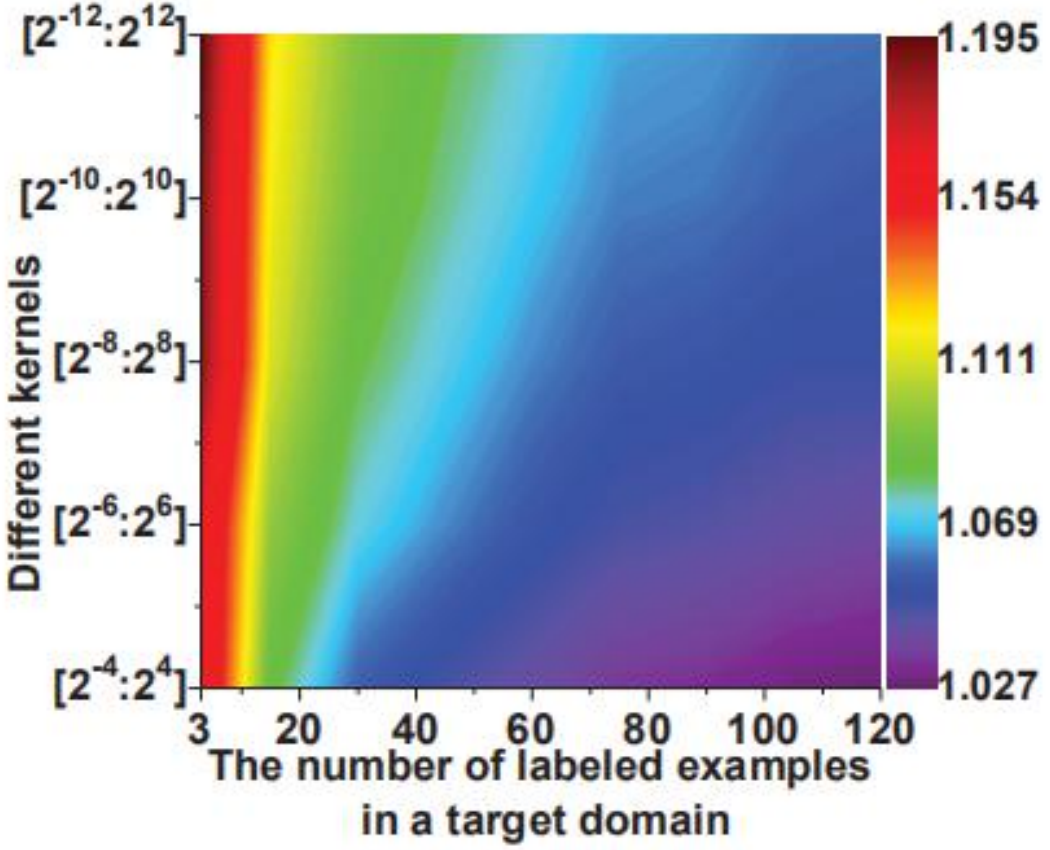
(e) bridge / helicopter / tripod → key / parrot / traffic-light



(f) caculator / straw / french-horn → doorknob / palm-tree / scissors

Experiments





Experiments