A Mathematical Framework for Quantifying Transferability in Multi-source Transfer Learning

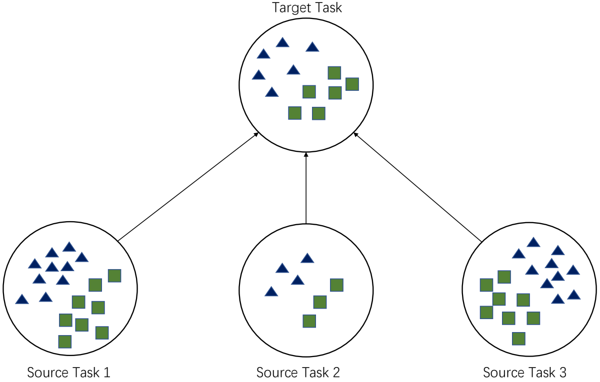
Xinyi Tong*, Xiangxiang Xu†, Shao-Lun Huang*, Lizhong Zheng†

*Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
†Massachusetts Institute of Technology

Tuesday 19th October, 2021

Motivation

Which source task is more helpful for the target task?



Motivation

We can intuitively claim that source task 1 helps more, i.e.,

Source task 1 has higher transferability.

Intuitively, the reasons are

- Source task 2: less samples $\rightarrow$ Sample size is important
- Source task 3: not like target task $\rightarrow$ Similarity is important

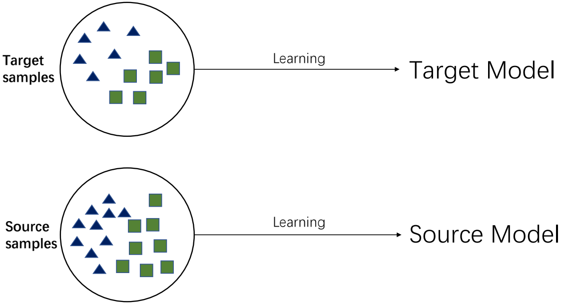
Motivation

What we hope to answer in this paper:

- Establish a mathematical framework to analyze transferability.
- Interpret all these factors.
- Apply the theoretical analyses to practical tasks.

Problem Formulation

First, how to establish the framework for transfer learning?



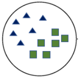
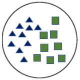
The final model using the source knowledge is **Optimal α?**

$$(1-\alpha)\cdot \text{Target Model} + \alpha \cdot \text{Source Model}$$

TBSI
Tsinghua-Berkeley Shenzhen Institute

Discrete Case

Sample $x \in \mathcal{X}$ & Label $y \in \mathcal{Y}$
We hope to learn the target distribution $P_{XY}^{(0)}$

- Target samples $\{(x_\ell^{(0)}, y_\ell^{(0)})\}_{\ell=1}^{n_0}$  i.i.d from $P_{XY}^{(0)} \to \hat{P}_{XY}^{(0)}$
- Source samples $\{(x_\ell^{(1)}, y_\ell^{(1)})\}_{\ell=1}^{n_1}$  i.i.d from $P_{XY}^{(1)} \to \hat{P}_{XY}^{(1)}$

We use $(1-\alpha)\hat{P}_{XY}^{(0)} + \alpha\,\hat{P}_{XY}^{(1)}$ to estimate $P_{XY}^{(0)}$.

Testing loss

How to evaluate the model? By the test data!

$$L_{\text{test}}^{(\alpha)} \triangleq \mathbb{E}\left[\chi^2\left(P_{XY}^{(0)}, (1-\alpha)\hat{P}_{XY}^{(0)} + \alpha\,\hat{P}_{XY}^{(1)}\right)\right] \tag{1}$$

Why not Log-loss?

What is optimal coefficient? $\longrightarrow$ Transferability

$$\alpha^* = \arg\min_{\alpha}\; L_{\text{test}}^{(\alpha)} \tag{2}$$

Testing loss

Theorem

The testing loss can be computed as

$$L_{\text{test}}^{(\alpha)} = \alpha^2 \chi^2 \left(P_{XY}^{(0)}, P_{XY}^{(1)} \right) + \frac{(1-\alpha)^2}{n_0} V^{(0)} + \frac{\alpha^2}{n_1} V^{(1)}, \tag{3}$$

and the optimal α^* is

$$\alpha^* = \frac{\frac{1}{n_0} V^{(0)}}{\chi^2(P_{XY}^{(0)}, P_{XY}^{(1)}) + \frac{1}{n_0} V^{(0)} + \frac{1}{n_1} V^{(1)}}, \tag{4}$$

where $V^{(0)} = |\mathcal{X}||\mathcal{Y}| - 1$ and $V^{(1)} = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \frac{P_{XY}^{(1)}(x,y)\left(1 - P_{XY}^{(1)}(x,y)\right)}{P_{XY}^{(0)}(x,y)}$.

Transferability

We claim that α^* is our transferability measure.

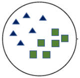
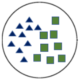
The affecting factors

- $\chi^2\left(P_{XY}^{(0)}, P_{XY}^{(1)}\right) \rightarrow$ distance!
- n_0 & $n_1 \rightarrow$ sample size!
- $|\mathcal{X}||\mathcal{Y}| - 1 \rightarrow$ task complexity!

Consistent with our intuition

Multi-source Transfer Learning

We can extend this to the multi-source case.

- Target task:  i.i.d from $P_{XY}^{(0)} \rightarrow \hat{P}_{XY}^{(0)}$
- Source task 1:  i.i.d from $P_{XY}^{(1)} \rightarrow \hat{P}_{XY}^{(1)}$
- $\cdots$
- Source task k:  i.i.d from $P_{XY}^{(k)} \rightarrow \hat{P}_{XY}^{(k)}$

We use $\alpha_0\hat{P}_{XY}^{(0)} + \alpha_1\hat{P}_{XY}^{(1)} + \cdots + \alpha_k\hat{P}_{XY}^{(k)}$ to estimate $P_{XY}^{(0)}$.

We have the testing loss

$$L_{\text{test}} = \chi^2\left(P_{XY}^{(0)}, \sum_{i=0}^{k}\alpha_i P_{XY}^{(i)}\right) + \sum_{i=0}^{k}\frac{\alpha_i^2}{n_i}V^{(i)}, \quad (5)$$

and we can find the optimal coefficients.

*We request $\sum_{i=0}^{k}\alpha_i = 1$ here.

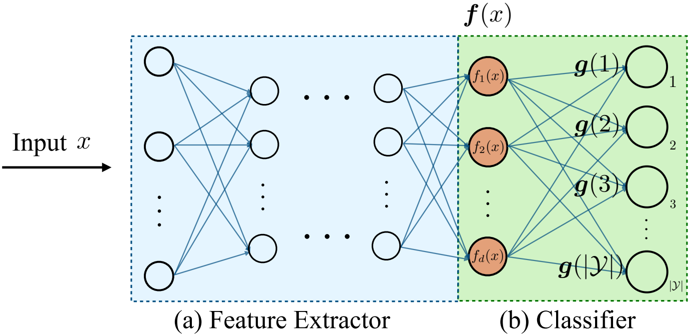
Continuous Case

Is our theory practical? Yes

However, there are 2 things we need to solve

- How the neural network models the distribution
- How to avoid the high dimensionality $|\mathcal{X}||\mathcal{Y}|$

Parametric Model



Discriminative Model

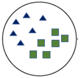
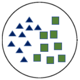
$$\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}(y|x) \triangleq P_Y^{(0)}(y)\left(1 + \boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y)\right), \tag{6}$$

Parametric Model

When f is fixed, we can train a classier g by samples under the referenced χ^2-loss.

$$\hat{\boldsymbol{g}}_i = \arg\min_{\boldsymbol{g}} \chi^2_R(\hat{P}^{(i)}_{XY}, P^{(0)}_X \tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X}),$$

where $\chi^2_R(P, Q) \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{(P(x,y)-Q(x,y))^2}{P^{(0)}_X(x)P^{(0)}_Y(y)}$.

- Target task:  $\rightarrow \hat{\boldsymbol{g}}_0 \quad (\mathbb{E}[\hat{\boldsymbol{g}}_0] = \boldsymbol{g}_0)$
- Source task 1:  $\rightarrow \hat{\boldsymbol{g}}_1 \quad (\mathbb{E}[\hat{\boldsymbol{g}}_1] = \boldsymbol{g}_1)$
- $\cdots$
- Source task k:  $\rightarrow \hat{\boldsymbol{g}}_k \quad (\mathbb{E}[\hat{\boldsymbol{g}}_k] = \boldsymbol{g}_k)$

Parametric Model

We use

$$\alpha_0 \tilde{P}_{Y|X}^{(\boldsymbol{f}, \hat{\boldsymbol{g}}_0)} + \alpha_1 \tilde{P}_{Y|X}^{(\boldsymbol{f}, \hat{\boldsymbol{g}}_1)} + \cdots + \alpha_k \tilde{P}_{Y|X}^{(\boldsymbol{f}, \hat{\boldsymbol{g}}_k)} \tag{7}$$

as our estimation. We have the testing loss

$$L_{\text{test}} = \chi_R^2 \left(P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f}, \boldsymbol{g}_0)}, \sum_{i=0}^{k} \alpha_i P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f}, \boldsymbol{g}_i)} \right) + \sum_{i=0}^{k} \frac{\alpha_i^2}{n_i} \tilde{V}^{(i)} + \chi_R^2 \left(P_{XY}^{(0)}, P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f}, \boldsymbol{g}_0)} \right)$$

Consistent with the theory in the discrete case

Factors

Transferability

$$(\alpha_0^*, \alpha_1^*, \cdots, \alpha_k^*) = \underset{(\alpha_0^*, \alpha_1^*, \cdots, \alpha_k^*): \sum_{i=0}^{k} \alpha_i = 1}{\arg\min} L_{\text{test}} \tag{8}$$

Let's see what's the affecting factors in the parametric model

- $\chi_R^2 \left(P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f}, \boldsymbol{g}_0)}, \sum_{i=0}^{k} \alpha_i P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f}, \boldsymbol{g}_i)} \right) \rightarrow$ distance!
- $n_i \rightarrow$ sample size!
- $\tilde{V}^{(i)} \rightarrow$ task complexity!

TBSI
Tsinghua-Berkeley Shenzhen Institute

Algorithm and Experimental Results

We have an iterative algorithm:

- $(\boldsymbol{f}, \boldsymbol{g}) \longleftarrow$ Training Loss with given $\alpha_0, \alpha_1, \cdots, \alpha_k$
- $(\alpha_0, \alpha_1, \cdots, \alpha_k) \longleftarrow$ Testing Loss with given $\boldsymbol{f}, \boldsymbol{g}$
- Until Converge

We made experiments on CIFAR-10, Office-31 and Office-Caltech datasets.

Conclusion

What's the contributions of our work?

- A theoretical analysis for transferability covering distance, sample sizes, task complexity at the same time
- An extension to continuous data
- A consistent algorithm that works