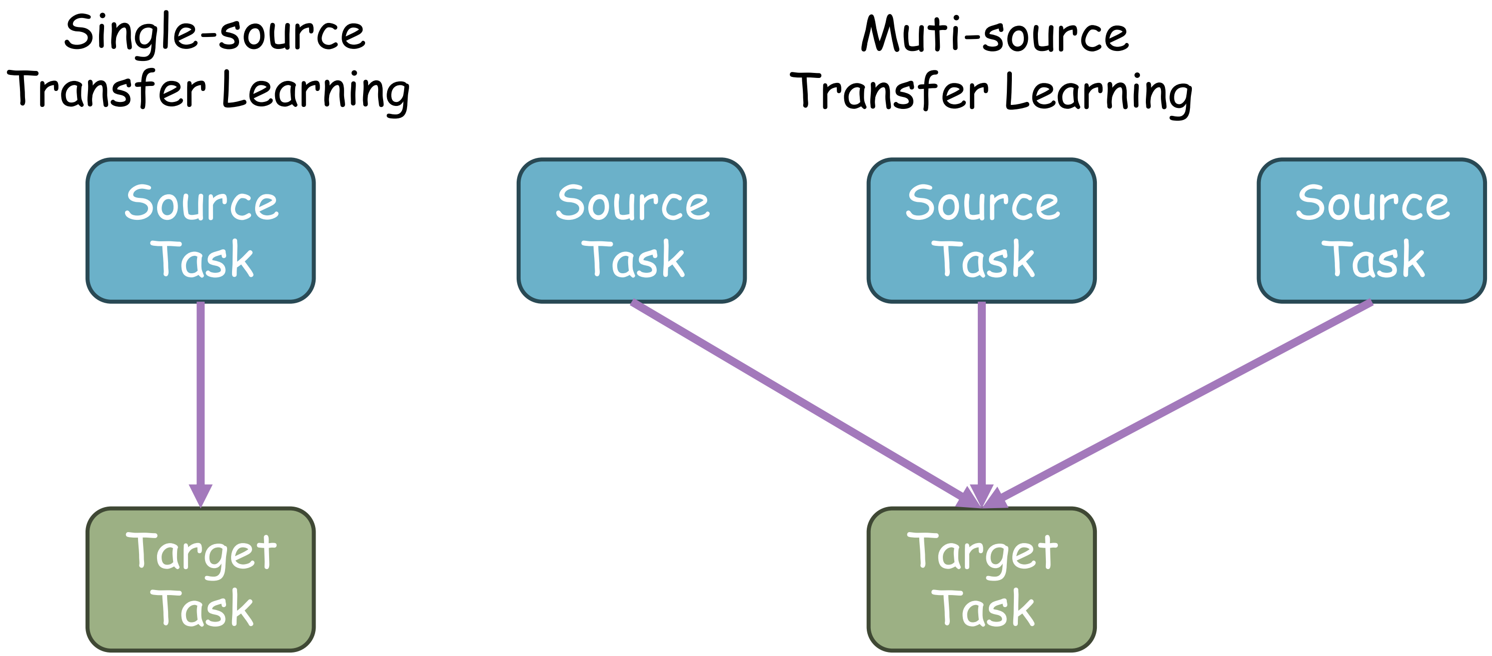
A High-Dimensional Statistical Method for Optimizing Transfer Quantities in Multi-Source Transfer Learning

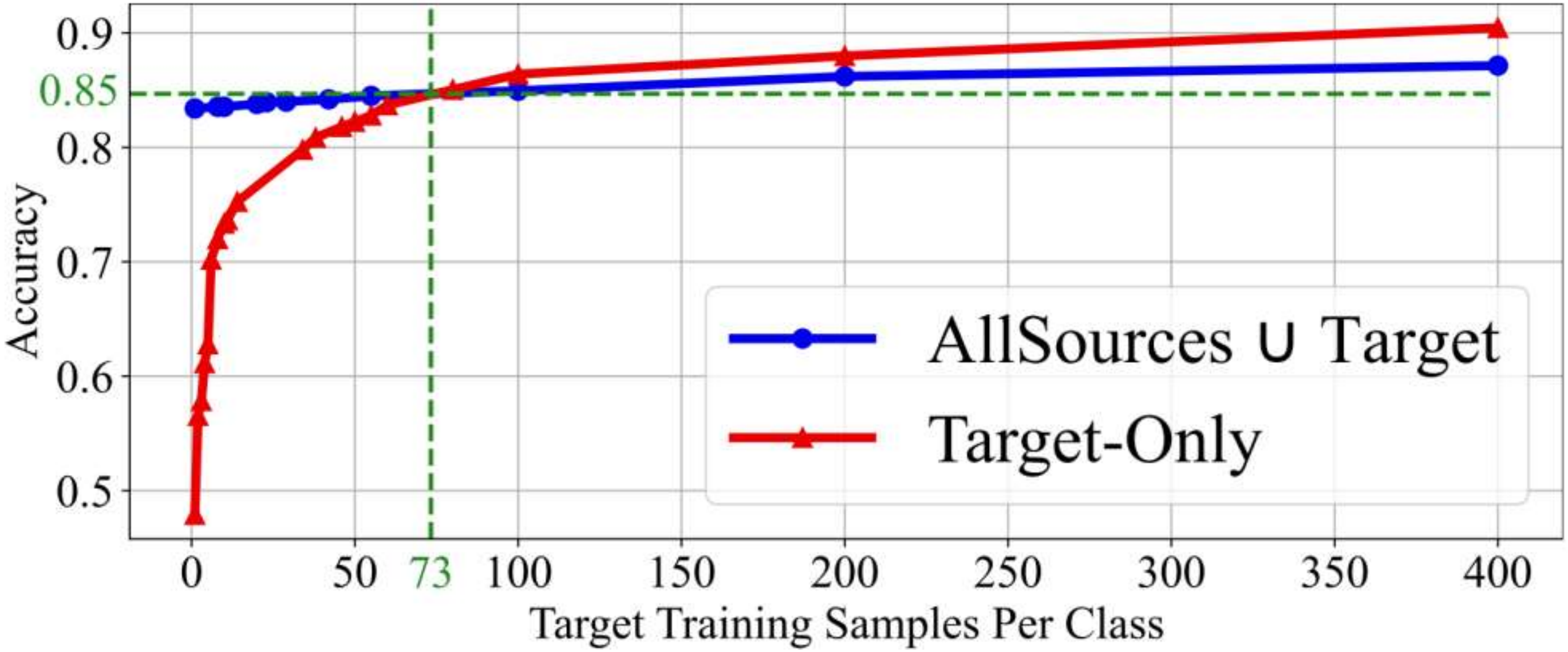
Qingyue Zhang*, Haohao Fu*, Guanbo Huang*, Yaoyuan Liang, Chang Chu, Tianren Peng, Yanru Wu, Qi Li, Yang Li, Shao-Lun Huang
Tsinghua University
zhangqy23@mails.tsinghua.edu.cn

Contents

Background: Transfer Learning



Background: Transfer Learning



What is the optimal transfer quantity of each source task?

Contents

Preliminaries: Asymptotic Normality of the MLE

Tsinghua University

- The maximum likelihood estimator (MLE) is defined as the maximizer of the empirical log-likelihood:

$$\hat{\underline{\theta}}_{\mathrm{MLE}} = \arg\max_{\underline{\theta}} \frac{1}{n} \sum_{x \in \mathcal{D}} \log P_{X;\underline{\theta}}(x)$$

- **Asymptotic Normality:** As the sample size increases, the distribution of the normalized estimation error converges in law to a multivariate Gaussian distribution

$$\sqrt{n}\left(\hat{\underline{\theta}}_{\mathrm{MLE}} - \underline{\theta}^*\right) \xrightarrow{d} \mathcal{N}(0, J(\underline{\theta}^*)^{-1})$$

- **The Fisher information matrix,** which characterizes the amount of information carried by the distribution about the parameter, is defined as

$$J(\underline{\theta})^{d\times d} = \mathbb{E}\left[\left(\frac{\partial}{\partial \underline{\theta}} \log P_{X;\underline{\theta}}\right)\left(\frac{\partial}{\partial \underline{\theta}} \log P_{X;\underline{\theta}}\right)^T\right].$$

Preliminaries: Problem Formulation

- **Target task:** Training samples $X^{N_0} = \{x_j\}_{j=1}^{N_0} \overset{\text{i.i.d.}}{\sim} P_{X;\underline{\theta}_0}$.
- **Source tasks:** Each source task $\mathcal{S}_i$ has $X_i^{N_i} = \{x_{i,j}\}_{j=1}^{N_i} \overset{\text{i.i.d.}}{\sim} P_{X;\underline{\theta}_i}$, $i \in [1, K]$.
- **Estimator:** The training process is formulated as a parameter estimation problem. The MLE $\hat{\underline{\theta}}$ is obtained using all target samples and a selected subset of source samples:

$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta}} \left[\sum_{x \in X^{N_0}} \log P_{X;\underline{\theta}}(x) + \sum_{i=1}^{K} \sum_{x \in X^{n_i}} \log P_{X;\underline{\theta}}(x) \right].$$

- **Objective:** Find optimal transfer quantities $n_1^*, \ldots, n_K^*$ by minimizing the expected K-L divergence between the true target distribution $P_{X;\underline{\theta}_0}$ and the learned one $P_{X;\hat{\underline{\theta}}}$.

$$n_1^*, \ldots, n_K^* = \arg\min_{n_1,\ldots,n_K} \mathbb{E}[D(P_{X;\underline{\theta}_0} || P_{X;\hat{\underline{\theta}}})].$$

Contents

Theorem (Theorem 1. Single-source transfer with 1-dimensional models)

In transfer between $P_{X;\theta_0}$ *and* $P_{X;\theta_1}$, *where* $\theta_0, \theta_1 \in \mathbb{R}$ *and* $|\theta_0 - \theta_1| = O(\frac{1}{\sqrt{N_0}})$. *Then, the K-L measure* $\mathbb{E}[D(P_{X;\underline{\theta}_0} || P_{X;\underline{\hat{\theta}}})]$ *can be expressed as:*

$$\frac{1}{2}\Bigg(\underbrace{\frac{1}{N_0 + n_1}}_{\text{variance term}} + \underbrace{\frac{n_1^2}{(N_0 + n_1)^2} t}_{\text{bias term}} \Bigg) + o\left(\frac{1}{N_0}\right), \text{ where } t \triangleq J(\theta_0)(\theta_1 - \theta_0)^2.$$

Optimal Transfer Quantity: *The optimal transfer quantity* n_1^* *is*

$$n_1^* = \begin{cases} N_1, & \text{if } N_0 \cdot t \leq 0.5 \\ \min\left(N_1, \frac{N_0}{2N_0 t - 1}\right), & \text{if } N_0 \cdot t > 0.5 \end{cases}. \tag{1}$$

Main Results

清華大学 Tsinghua University

Theorem (Theorem 2. Single-source transfer with high-dimensional models)

In transfer between $P_{X;\theta_0}$ and $P_{X;\theta_1}$, where $\theta_0, \theta_1 \in \mathbb{R}^d$ and $||\theta_0 - \theta_1|| = O(\frac{1}{\sqrt{N_0}})$. Then, the K-L measure $\mathbb{E}[D(P_{X;\underline{\theta}_0}||P_{X;\hat{\underline{\theta}}})]$ can be expressed as:

$$\frac{d}{2}\left(\frac{1}{N_0+n_1}+\frac{n_1^2}{(N_0+n_1)^2}t\right)+o\left(\frac{1}{N_0}\right), \text{where } t \triangleq \frac{(\underline{\theta}_1-\underline{\theta}_0)^\top J(\underline{\theta}_0)(\underline{\theta}_1-\underline{\theta}_0)}{d}.$$

Optimal Transfer Quantity: *The optimal transfer quantity n_1^* is*

$$n_1^* = \begin{cases} N_1, & \text{if } N_0 \cdot t \leq 0.5 \\ \min\left(N_1, \frac{N_0}{2N_0t-1}\right), & \text{if } N_0 \cdot t > 0.5 \end{cases}. \tag{2}$$

Main Results

Theorem (Theorem 3. Multi-source transfer with high-dimensional models)

In the multi-source setting, we denote $s = \sum_{i=1}^{K} n_i$, *and* $\alpha_i = \frac{n_i}{s}$. *Then, the K-L measure can be expressed as:*

$$\frac{d}{2}\left(\frac{N_0}{(N_0+s)^2} + \frac{s^2}{(N_0+s)^2}t\right) + o\left(\frac{1}{N_0}\right), t = \frac{\underline{\alpha}^T \Theta^T J(\underline{\theta}_0)\Theta\underline{\alpha}}{d},$$

where $\underline{\alpha} = [\alpha_1, \ldots, \alpha_K]^\top$, *and* $\Theta^{d\times K} = [\underline{\theta}_1 - \underline{\theta}_0, \ldots, \underline{\theta}_K - \underline{\theta}_0]$.

Optimal Transfer Quantities: *The optimal quantities* n_i^* *are obtained by minimizing the K-L measure. Specificially, we perform a grid search over the feasible range of* s, *and for each candidate* s', *solve a quadratic programming problem to obtain the optimal* $\underline{\alpha}'$ *under the constraint set* $\mathcal{A}(s')$. *The final optimal pair* $(s^*, \underline{\alpha}^*)$ *is selected as the one yielding the minimum value of the objective function among all candidate pairs. Finally, we use* $n_i^* = s^*\alpha_i^*$.

Main Results: Algorithm

Algorithm 1 OTQMS: Training

Input: Target data $D_{\mathcal{T}} = \{(z_{\mathcal{T}}^i, y_{\mathcal{T}}^i)\}_{i=1}^{N_0}$, source data $\{D_{S_k} = \{(z_{S_k}^i, y_{S_k}^i)\}_{i=1}^{N_k}\}_{k=1}^{K}$, model type $f_{\underline{\theta}}$ and its parameters $\underline{\theta}_0$ for target task and $\{\underline{\theta}_k\}_{k=1}^{K}$ for source tasks, parameter dimension d. // z represents the feature and y represents the label

Parameter: Learning rate η.

Initialize: randomly initialize $\underline{\theta}_0$, use parameters of pretrained source models to initialize $\{\underline{\theta}_k\}_{k=1}^{K}$.

Output: a well-trained $\underline{\theta}_0$ for target task model $f_{\underline{\theta}_0}$.

$D_{train} \leftarrow D_{\mathcal{T}}$ // Initialize the trainning dataset by target task samples

repeat // Use dynamic strategy to train the target task

$\mathcal{L}_{train} \leftarrow \frac{1}{|D_{train}|} \sum_{(y^i, z^i) \in D_{train}} \ell\left(y^i, f_{\underline{\theta}_0}(z^i)\right)$

$\underline{\theta}_0 \leftarrow \underline{\theta}_0 - \eta \nabla_{\underline{\theta}_0} \mathcal{L}_{train}$

$\Theta \leftarrow [\underline{\theta}_1 - \underline{\theta}_0, \ldots, \underline{\theta}_K - \underline{\theta}_0]^T$

$J(\underline{\theta}_0) \leftarrow (\nabla_{\underline{\theta}_0} \mathcal{L}_{train})(\nabla_{\underline{\theta}_0} \mathcal{L}_{train})^T$

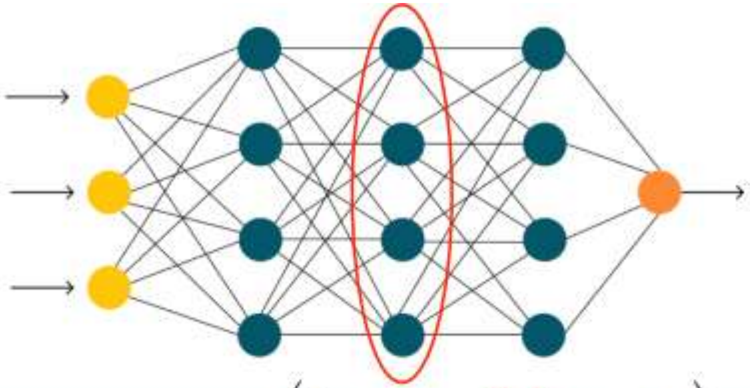
$(s^*, \underline{\alpha}^*) \leftarrow \arg\min_{(s, \underline{\alpha})} \frac{d}{2} \left(\frac{1}{N_0 + s} + \frac{s^2}{(N_0+s)^2} \frac{\underline{\alpha}^T \Theta^T J(\underline{\theta}_0) \Theta \underline{\alpha}}{d} \right)$

$D_{source} \leftarrow \bigcup_{k=1}^{K} \left\{ D_{S_k}^* \middle| D_{S_k}^* \overset{\text{rand}}{\subseteq} D_{S_k}, |D_{S_k}^*| = s^* \alpha_k^* \right\}$

$D_{train} \leftarrow D_{source} \bigcup D_{\mathcal{T}}$ // Update the trainning dataset

until $\underline{\theta}_0$ converges;

Dynamic Strategy



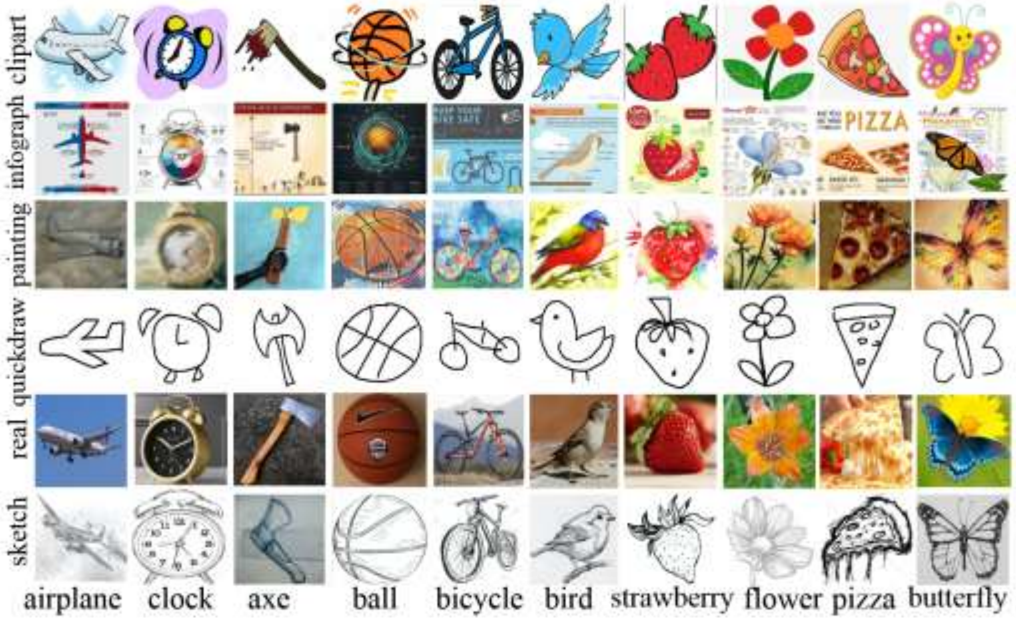
$$\theta_{t+1} = \theta_t + \eta \cdot \left(\sum_{i=1}^{n} \nabla_{\theta_t} \ell(z_i) + \sum_{j=1}^{M} \sum_{i=1}^{k_i} \nabla_{\theta_t} \ell(z_i^{(j)}) \right)$$

- The quantities in computing the optimal number of samples can be estimated from current parameters;
- Update the parameters in SGD algorithm with both target samples and k_i samples from source task i;
- Iteratively update the optimal sample numbers and network parameters.

Main Results: Dataset of Experiment

Office-Home

Four domains, each formulated as a 65-class classification task.



Domainnet

Six domains, each formulated as a 345-class classification task.

Main Results: Experiment

Table 2: **Multi-Source Transfer Performance on DomainNet and Office-Home.** The arrows indicate transfering from the rest tasks. The highest/second-highest accuracy is marked in Bold/Underscore form respectively.

Method	Backbone	DomainNet							Office-Home				
		→C	→I	→P	→Q	→R	→S	Avg	→Ar	→Cl	→Pr	→Rw	Avg
Unsupervised-all-shots													
MSFDA[19]	ResNet50	66.5	21.6	56.7	20.4	70.5	54.4	48.4	75.6	62.8	84.8	<u>85.3</u>	77.1
DATE[9]	ResNet50	-	-	-	-	-	-	-	75.2	60.9	**85.2**	84.0	76.3
M3SDA[17]	ResNet101	57.2	24.2	51.6	5.2	61.6	49.6	41.5	-	-	-	-	-
Supervised-10-shots													
Few-Shot Methods:													
H-ensemble[29]	ViT-S	53.4	21.3	54.4	19.0	70.4	44.0	43.8	71.8	47.5	77.6	79.1	69.0
MADA[30]	ViT-S	51.0	12.8	60.3	15.0	**81.4**	22.7	40.5	**78.4**	58.3	82.3	85.2	76.1
MADA[30]	ResNet50	66.1	23.9	<u>60.4</u>	<u>31.9</u>	<u>75.4</u>	52.5	51.7	72.2	<u>64.4</u>	82.9	81.9	75.4
MCW[12]	ViT-S	54.9	21.0	53.6	20.4	70.8	42.4	43.9	68.9	48.0	77.4	**86.0**	70.1
WADN[20]	ViT-S	68.0	29.7	59.1	16.8	74.2	55.1	50.5	60.3	39.7	66.2	68.7	58.7
Source-Ablation Methods:													
Target-Only	ViT-S	14.2	3.3	23.2	7.2	41.4	10.6	16.7	40.0	33.3	54.9	52.6	45.2
Single-Source-Avg	ViT-S	50.4	22.1	44.9	24.7	58.8	42.5	40.6	65.2	53.3	74.4	72.7	66.4
Single-Source-Best	ViT-S	60.2	28.0	55.4	28.4	66.0	49.7	48.0	72.9	60.9	80.7	74.8	72.3
AllSources ∪ Target	ViT-S	<u>71.7</u>	<u>32.4</u>	60.0	31.4	71.7	<u>58.5</u>	<u>54.3</u>	77.0	62.3	84.9	84.5	<u>77.2</u>
OTQMS (Ours)	ViT-S	**72.8**	**33.8**	**61.2**	**33.8**	73.2	**59.8**	**55.8**	<u>78.1</u>	**64.5**	**85.2**	84.9	**78.2**

Main Results: Experiment

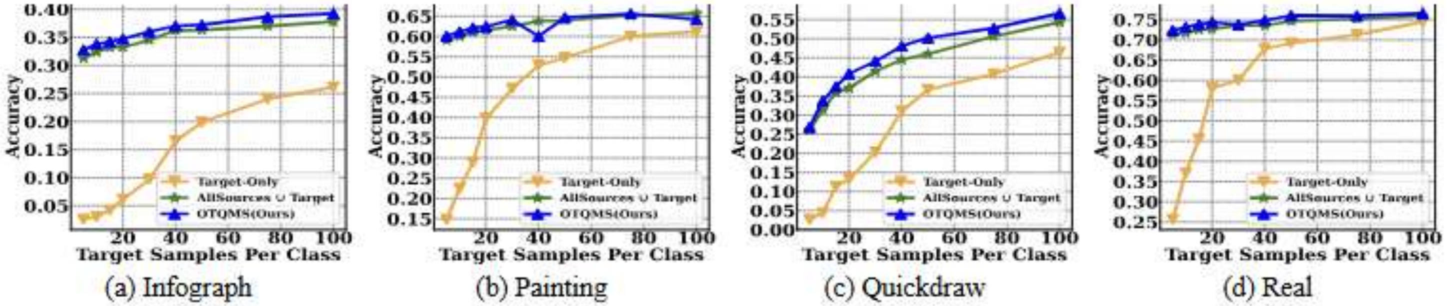


Figure 3: Performance comparison with increasing target shots up to 100 per class on DomainNet dataset (I, P, Q and R domains). OTQMS (blue) outperforms other methods.

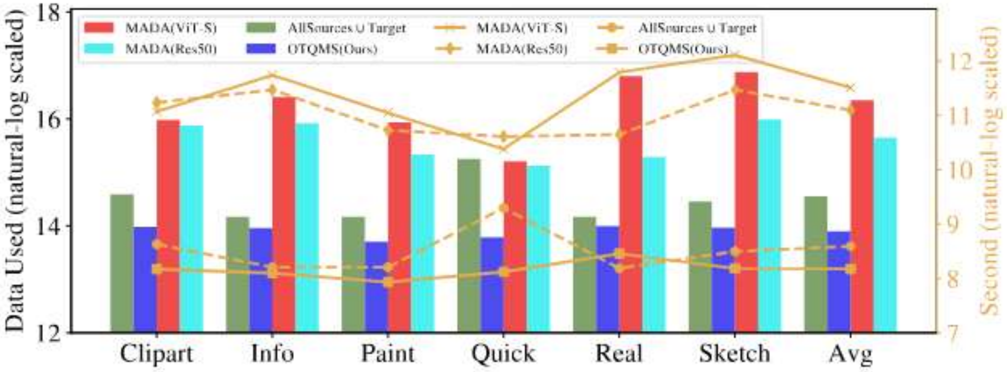


Figure 4: Data efficiency comparison of average sample usage and training time on DomainNet dataset, the left vertical axis represents the amount of sample usage, with green bars indicating AllSources ∪ Target data counts, blue bars about OTQMS, red bars about MADA(ViT-S) and azury bars about MADA(Res50), while the right orange vertical axis and lines represent training time.

Main Results: Experiment

Tsinghua University

Table 6: Multi-Source Transfer with LoRA on Office-Home. We apply LoRA on ViT-B backbone for PEFT.

Method	Backbone	Office-Home				
		→Ar	→Cl	→Pr	→Rw	Avg
Supervised-10-shots Source-Ablation:						
Target-Only	ViT-B	59.8	42.2	69.5	72.0	60.9
Single-Source-avg	ViT-B	72.2	59.9	82.6	81.0	73.9
Single-Source-best	ViT-B	74.4	61.8	84.9	81.9	75.8
AllSources ∪ Target	ViT-B	81.1	66.0	88.0	89.2	81.1
OTQMS (Ours)	ViT-B	**81.5**	**68.0**	**89.2**	**90.3**	**82.3**

Table 4: Multi-task performance on four tasks of Office-Home.

Method	Backbone	Office-Home				
		Ar	Cl	Pr	Rw	Avg
Single-task	ViT-S	66.7	62.3	87.8	68.6	71.4
OTQMS	ViT-S	**81.7**	**76.0**	**88.6**	**87.5**	**83.5**

Contents

Conclusion

- **Theoretical Framework:** We formulate multi-source transfer learning as a parameter estimation problem and derive solutions for the **optimal transfer quantity** of each source by minimizing a K-L divergence–based generalization error in the asymptotic regime.
- **Algorithm Design:** We develop **OTQMS**, a dynamic and data-efficient algorithm that iteratively updates transfer quantities using empirical Fisher information, enabling adaptive resampling and improved target model training.
- **Experimental Results:** Experiments on **DomainNet** and **Office-Home** demonstrate that OTQMS achieves higher accuracy and better data efficiency than state-of-the-art methods, and remains robust under different shot settings and architectures.

NEURAL INFORMATION PROCESSING SYSTEMS

清華大學 Tsinghua University

Thanks for your attention !

https://github.com/zqy0126/OTQMS