Supplementary Material
OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations

Yang Tan, Yang Li$^{\boxtimes}$, Shao-Lun Huang
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
tany19@mails.tsinghua.edu.cn, {yangli, shaolun.huang}@sz.tsinghua.edu.cn

1. Visualization of Datasets

Some data instances of DomainNet [3] and Office31 [4] are depicted in Figure 1, 2 respectively. We can see that the five domains of DomainNet have unique image styles and the three domains of Office31 differ in their image acquisition devices and environments. We can also observe that the learning difficulty of images in each domain differs widely. For instance, *Quickdraw* has the least visual complexity among DomainNet's five domains for just having black-and-white lines; Product photos in *Amazon* have far less uncertainty than those in other datasets that contain real world photos for having clean background and simple lighting condition. Such difference is reflected in our transferability analysis using OTCE.

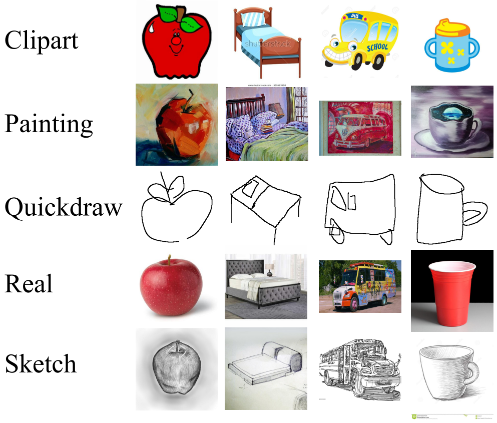

Figure 1. Visualization of data instances from DomainNet [3].

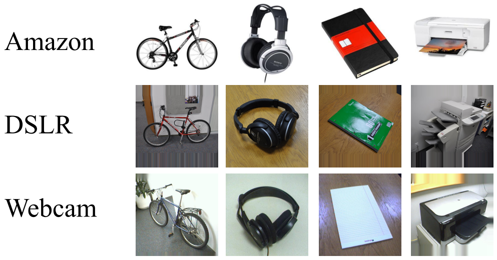

Figure 2. Visualization of data instances from Office31 [4].

2. Analysis of OTCE

Recall that our OTCE score characterizes transferability based on two factors: *domain difference* W_D measured by cross-domain Wasserstein distance and *task difference* W_T measured by the conditional entropy between the source and target labels under the optimal coupling between two domains. In this section, we examine the relationship between each factor and transferability under the three transfer settings discussed in the paper.

First, observing the Wasserstein distances of all transfer instances (left column in Figure 3), we find that the domain difference is relatively stable among different target tasks when source and target domains are fixed. This is reasonable since domain difference should be task agnostic. Meanwhile, domain difference shows a generally negative correlation with the transfer accuracy, which is most obvious when the category set size is fixed (row 5-6 of Figure 3). The only exception in this case is when the target domain is *Quickdraw* (represented by red points). Due to the low visual complexity in line drawings, most features can achieve relatively high classification accuracy on *Quickdraw* despite being trained on very different source domains. Note that we do not require that the correlation of domain difference and transfer accuracy to be strictly negative. Because in our unified framework, domain difference and task difference are coupled due to the coupling matrix computed via OT.

Second, as we emphasize in the paper (Section 3.3) that task difference W_T measured by conditional entropy alone is not sufficient to characterize cross-domain transferability, we present the experimental evidence to support this finding by looking at the correlation between W_T and the transfer accuracy. In Figure 3, our OTCE score (right column) shows significantly higher correlation with the transfer accuracy compared to conditional entropy (middle column). The improvement is most notable under the few-shot and fixed category set size setting, which shows that incorporating domain difference W_D can indeed improve the robustness of transferability prediction in weakly supervised and challenging scenarios.

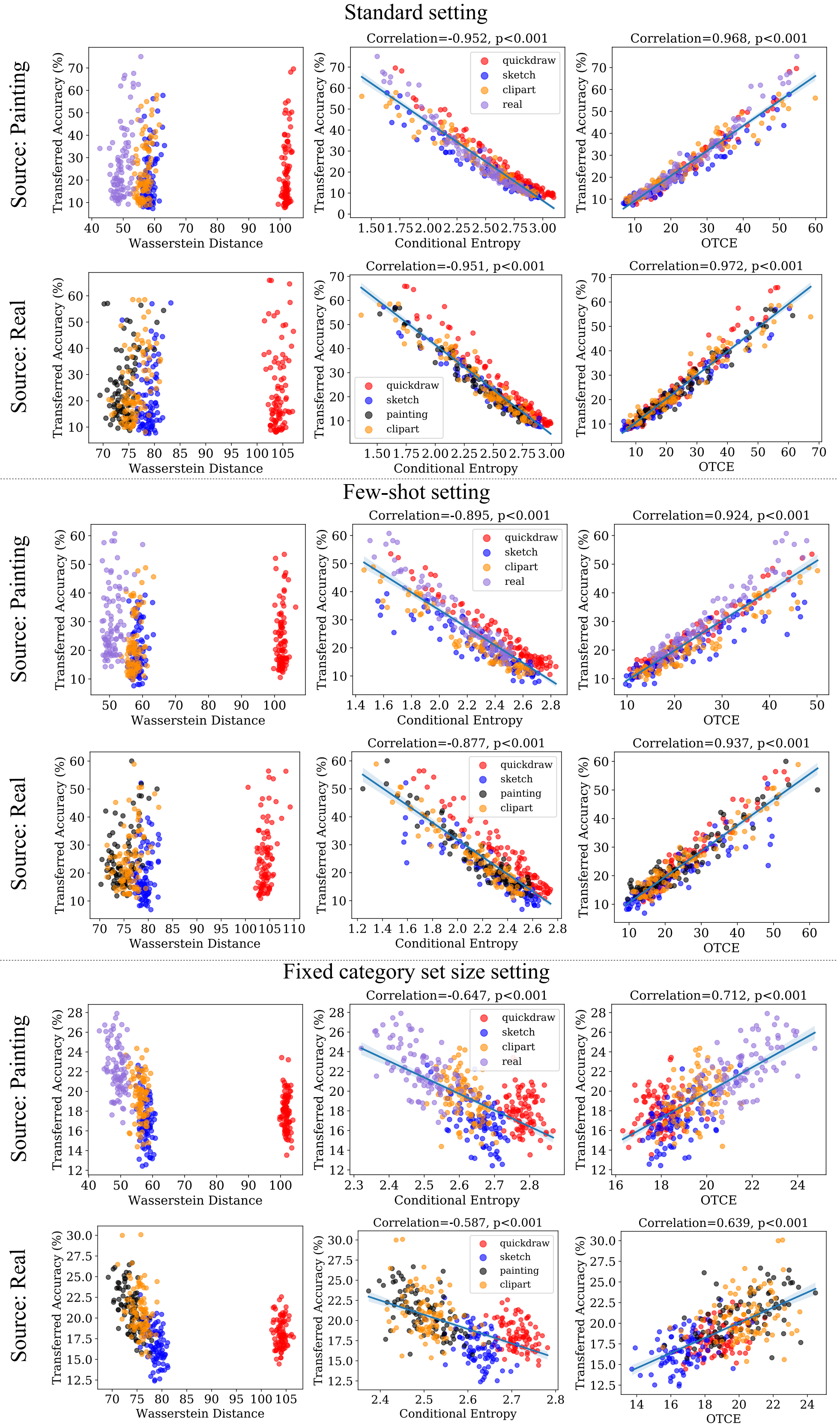

Figure 3. Visualization of OTCE score and its components, i.e., Wasserstein distance (domain difference W_D) and Conditional Entropy (task difference W_T). Points represent different target tasks.

Table 1. Quantitative comparisons evaluated by Pearson correlation coefficients of transferability metrics and transfer accuracy through Retrain head under cross-domain cross-task transfer settings. Superscript * denotes $p > 0.001$.

Dataset	Experimental setting			**OTCE**	OT-based NCE	LEEP[2]	NCE[5]	H-score[1]
	Source domain	Target domain	Data property					
DomainNet	C	P,Q,R,S	standard	**0.969**	**0.960**	0.919	0.787	0.864
	P	C,Q,R,S	standard	**0.968**	**0.952**	0.886	0.812	0.858
	Q	C,P,R,S	standard	**0.963**	**0.963**	0.942	0.935	0.843
	R	C,P,Q,S	standard	**0.972**	**0.951**	0.892	0.851	0.870
	S	C,P,Q,R	standard	**0.960**	**0.959**	0.952	0.954	0.882
Office31	A	D,W	standard	**0.829**	**0.813**	0.805	0.796	0.590
	D	A,W	standard	**0.880**	0.843	**0.857**	0.849	0.441
	W	A,D	standard	**0.863**	0.803	**0.811**	0.804	0.489
			average	**0.926**	**0.906**	0.883	0.849	0.730
DomainNet	C	P,Q,R,S	fixed category set size	**0.757**	**0.729**	0.614	0.535	-0.599
	P	C,Q,R,S	fixed category set size	**0.712**	**0.647**	0.480	0.418	-0.541
	Q	C,P,R,S	fixed category set size	**0.352**	**0.306**	0.213	0.269	-0.288
	R	C,P,Q,S	fixed category set size	**0.639**	**0.587**	0.465	0.440	-0.100^*
	S	C,P,Q,R	fixed category set size	**0.435**	**0.443**	0.381	0.427	-0.302
			average	**0.579**	**0.542**	0.431	0.418	-0.346
DomainNet	C	P,Q,R,S	few-shot	**0.920**	**0.907**	0.843	0.713	0.767
	P	C,Q,R,S	few-shot	**0.924**	**0.895**	0.812	0.737	0.807
	Q	C,P,R,S	few-shot	**0.852**	**0.857**	0.836	0.825	0.786
	R	C,P,Q,S	few-shot	**0.937**	**0.877**	0.787	0.744	0.814
	S	C,P,Q,R	few-shot	**0.922**	**0.901**	0.886	0.884	0.834
Office31	A	D,W	few-shot	**0.840**	**0.826**	0.803	0.793	0.640
	D	A,W	few-shot	**0.933**	**0.931**	0.923	0.930	0.413
	W	A,D	few-shot	**0.927**	**0.932**	0.920	0.926	0.277^*
			average	**0.907**	**0.891**	0.851	0.819	0.633

3. Study of OT-based NCE

Here we define an alternative transferability metric, **OT-based NCE** to be the negative conditional entropy ($-W_T$) mentioned earlier. Although it is not as robust as OTCE in estimating transferability, it does not require auxiliary task for parameter fitting, and thus is more efficient. We make quantitative comparisons of OT-based NCE with previous transferability metrics LEEP [2], NCE [5] and H-score [1]. Results shown in Table 1 demonstrate that our OT-based NCE also outperforms previous metrics on average. To conclude, our proposed two transferability metrics, i.e., OTCE and OT-based NCE, possess different advantages and readers can choose flexibly according to their need.

- **OTCE.** It suits the scenario which needs the most accurate transferability estimation or there are many target tasks. In addition, the learned coefficients of domain difference and task difference may benefit some downstream transfer learning applications.
- **OT-based NCE.** It is a simple implementation of OTCE, providing relatively coarse but more efficient transferability estimation without extra computation in auxiliary tasks. Although it is not as accurate as OTCE, it still averagely outperforms SOTA analytical metrics.

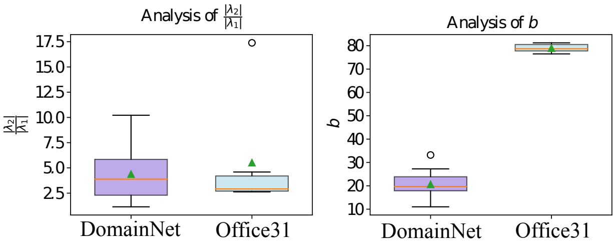

Figure 4. Analysis of learned coefficients.

4. Analysis on Learned Coefficients

Setting $\lambda_1 = 1, \lambda_2 = 0$ or $\lambda_1 = 0, \lambda_2 = 1$, i.e., only using domain difference or task difference (OT-based NCE) to characterize transferability, does not perform as good as OTCE, as depicted in the Figure 3 and Table 1. Moreover, we further analyze the learned coefficients in all experimental settings. We found that bias b was stable for the same cross-domain dataset (shown in Figure 4), but λ_1, λ_2 among different transfer configurations varied irregularly. On one hand, the importance of domain difference and task difference varies for different cross-domain configurations. On the other hand, differences are calculated in the feature space of source model, so the learned coefficients have different scales among source models. Generally, task difference is more important in describing trans-

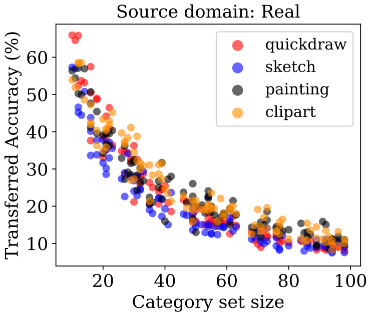

Figure 5. Visualization of transfer accuracy v.s. category set size (number of categories). Points represent different target tasks.

ferability ($\frac{|\lambda_2|}{|\lambda_1|} > 1$). To conclude, we recommend using auxiliary tasks to learn the coefficients for obtaining the most accurate transferability estimation for the given cross-domain transfer configuration.

5. Study of Category Set Size

Observations from our transferability experiments indicate that transfer accuracy drops when category set size (number of categories) increases, as shown in Figure 5. A larger category set generally makes it more difficult to learn the target task well under the same training setting. Such differences in the intrinsic complexity of the target task tends to overshadow the more subtle variations in transferability due to task and domain relatedness. Thus we make the quantitative comparisons under the fixed $category\_set\_size = 50$ setting (Section 4.3 in paper) to show that our OTCE score is capable of capturing these subtle variations.

References

[1] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2309–2313. IEEE, 2019. 3

[2] Cuong V Nguyen, Tal Hassner, Cedric Archambeau, and Matthias Seeger. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, 2020. 3

[3] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 1

[4] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 1

[5] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1405, 2019. 3