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

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

Abstract

Transfer learning across heterogeneous data distributions (a.k.a. domains) and distinct tasks is a more general and challenging problem than conventional transfer learning, where either domains or tasks are assumed to be the same. While neural network based feature transfer is widely used in transfer learning applications, finding the optimal transfer strategy still requires time-consuming experiments and domain knowledge. We propose a transferability metric called Optimal Transport based Conditional Entropy (OTCE), to analytically predict the transfer performance for supervised classification tasks in such cross-domain and cross-task feature transfer settings. Our OTCE score characterizes transferability as a combination of domain difference and task difference, and explicitly evaluates them from data in a unified framework. Specifically, we use optimal transport to estimate domain difference and the optimal coupling between source and target distributions, which is then used to derive the conditional entropy of the target task (task difference). Experiments on the largest cross-domain dataset DomainNet and Office31 demonstrate that OTCE shows an average of 21% *gain in the correlation with the ground truth transfer accuracy compared to state-of-the-art methods. We also investigate two applications of the OTCE score including source model selection and multi-source feature fusion.*

1. Introduction

Transfer learning is a useful learning paradigm to improve the performance on target tasks with the help of related source tasks (or source models), especially when only few labeled target data are available for supervision [30, 37]. A *Transferability* metric can quantitatively reveal how easy it is to transfer knowledge learned from a source task to the target task [13, 4, 36, 24]. It indeed provides a road map for conducting transfer learning in practice, e.g., selecting highly transferable tasks for joint training [41], or understanding task relationships for source model selection [5, 1, 38, 24].

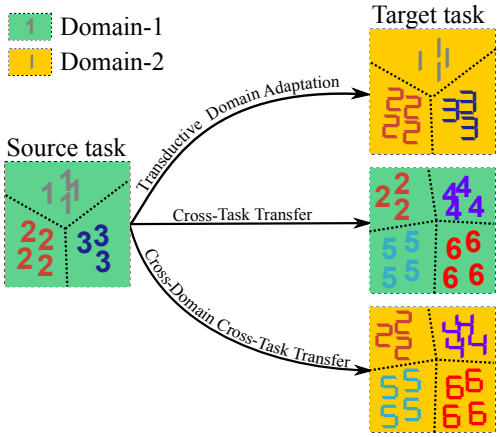

Figure 1. Illustration of three different transfer learning scenarios, i.e., transductive domain adaptation [27], cross-task transfer [5, 24] and the cross-domain cross-task transfer we investigating. We take the digital number classification as an example, where the cross-domain cross-task transfer setting suffers both domain difference and task difference.

While theoretical results in transfer learning such as [6, 8, 9, 21] suggest that task relationship can be modeled by certain divergence between the source and target data generating distributions, they are difficult to estimate in practice when target training data is limited. Previous transferability metrics [41, 40, 1] empirically calculate the task relationships indicated by training loss or validation accuracy, thus they need to retrain the source model involving expensive computation. Recent analytical metrics [5, 38, 24] are limited by strict assumptions on data. NCE [38] assumes that source and target tasks share the same input instances; H-score [5] assumes that source and target data are distributed in the same domain. Although LEEP [28] does not make any assumptions on source and target data except for having the same input size, it does not work sufficiently well under the cross-domain setting.

In this paper, we investigate the transferability estimation problem for classification tasks under the more challenging

✉ Corresponding author. This research is funded by Natural Science Foundation of China 62001266.

cross-domain cross-task setting, as illustrated in Figure 1. For most transfer learning problems we encounter in practice, we cannot assume the source and target data are generated from the same distribution (domain) since domain gaps commonly exist in real life due to different acquisition devices and different physical environments. Meanwhile, we also cannot always assume no task difference exists in a transfer learning application as in *transductive domain adaptation* [27], i.e., source and target tasks have the same category set. We emphasize that the *cross-domain cross-task* setting is more challenging compared to previous settings that require shared input data or same domain, since both *domain difference* and *task difference* deteriorate the transfer performance [27, 5, 26].

To this end, we propose a novel cross-domain cross-task transferability metric called the **O**ptimal **T**ransport based **C**onditional **E**ntropy, abbreviated as **OTCE** score. On one hand, compared to the empirical methods [41, 40, 1] that need to retrain the source model using gradient descent to estimate the empirical transfer error, our metric is more efficient (about 75x faster) to compute. On the other hand, our OTCE score explicitly learns the *domain difference* and *task difference* in a unified framework, providing a more interpretable result compared to recent analytical metrics [24, 38, 5].

More specifically, we measure the domain difference between source and target data using Wasserstein distance computed by solving the classic Optimal Transport (OT) [18, 29] problem. The OT problem also estimates the joint probability between source and target samples, which allows us to derive the task difference in terms of the conditional entropy between the source and target task labels. Finally, we learn a linear model of transfer accuracy on domain difference and task difference, drawing transfer experience evaluated on a few auxiliary tasks. Albeit its simplicity, the learned model makes it easier to decompose transferability into different factors through model coefficients.

Extensive experiments on the largest cross-domain dataset DomainNet [28] and Office31 [34] demonstrate that our OTCE score shows significantly higher correlation with transfer accuracy, i.e., predicting the transfer performance more accurately with an average of 21% gain compared to state-of-the-art metrics [24, 38, 5]. In addition, we further investigate two applications of transferability in source model selection and multi-source feature fusion. In summary, our contributions are follows:

1) To our knowledge, we are the first to analytically investigate the transferability estimation problem for supervised classification tasks under the more general and challenging cross-domain cross-task setting.

2) We propose a novel cross-domain cross-task transferability metric OTCE score which can explicitly evaluate *domain difference* and *task difference* in a unified framework, and predict the transfer performance in advance.

3) We show consistent superior performance in predicting transfer performance compared to state-of-the-art metrics and also investigate the applications of OTCE score in source model selection and multi-source feature fusion.

2. Related work

Our work is closely related to three fields in the transfer learning area [27, 20], i.e., empirical studies on transferability, analytical studies on transferability and task relatedness.

Empirical studies on transferability. Taskonomy [41] pioneers the investigation in empirically building a taxonomy of tasks. They retrain the source model on each target task and evaluate transfer performance to build up a non-parametric transferablity score called 'task affinity'. Task2Vec [1] embeds tasks into a low-dimensional vector space so that transferability can be measured using a non-symmetric distance metric. Task2Vec also needs to retrain the large-scale probe network on target task, and then compute the Fisher information matrix to obtain embedding vector. Ying *et al*. [39] propose to learn transfer skills from previous transfer learning experiences, and then apply such skills for future target tasks. Generally, these empirical methods involve expensive computation for training, which is mostly avoided in our approach.

Analytical studies on transferability. The advantage of analytical methods is computational efficiency. Previous H-score [5] is an information-theoretic approach for analytically evaluating transferability through solving a HGR maximum correlation problem. They focus on the *task transfer learning* problem, which assumes the same input domain among tasks. NCE [38] adopts conditional entropy to evaluate transferability and task hardness under a particular setting, i.e., source and target tasks share the same input instances but different labels. They provide a derivation that the empirical transferability is lower bounded by the negative conditional entropy. The recently proposed LEEP [24] score is a more general metric compared to the previous two methods. It is defined by the average log-likelihood of the expected empirical predictor, which predicts the dummy label distributions for target data in source label space and then compute the empirical conditional distribution of target label given the dummy source label. In general, these methods either have strict assumptions on data or do not work sufficiently well in a cross-domain setting.

Task relatedness. Although some theoretic analysis of generalization bounds [22, 6, 8, 7, 9, 21] in transfer learning and multi-task learning have shed insights on transferability estimation, it is difficult to verify whether their assumptions are satisfied in practical data and even more difficult to compute exactly. Meanwhile, such distance metrics including $\mathcal{F}$-relatedness [8], $\mathcal{A}$-distance [7] and discrepancy distance [21] are symmetric while the transferability metric

should be non-symmetric since transferring from one task to another is different from transferring in the reverse direction [24]. Besides, a recent work [3] using Optimal Transport (OT) to evaluate dataset distances also measures task relatedness to some extent. Task relatedness is also studied in multi-task learning, since weakly related tasks may worsen the performance compared to single task learning. [17, 31, 32] utilize some prior human knowledge to join highly related tasks. Other works [19, 23] are capable of dynamically adjusting the relatedness of tasks during the training phase.

3. Transferability Measure via OTCE

In this section, we start with presenting the transferability definition of classification tasks. Then we introduce some preliminary of Optimal Transport (OT) before detailing our proposed OTCE score.

3.1. Transferability Definition

Formally, suppose we have a pair of source and target classification tasks whose data are $D_s = \{(x_s^i, y_s^i)\}_{i=1}^m \sim P_s(x, y)$ and $D_t = \{(x_t^i, y_t^i)\}_{i=1}^n \sim P_t(x, y)$ respectively, where $x_s^i, x_t^i \in \mathcal{X}$ and $y_s^i \in \mathcal{Y}_s, y_t^i \in \mathcal{Y}_t$. Note that $x_s^i, x_t^i \in \mathcal{X}$ only implies source and target instances have the same input dimension, but they still reside in different domains, i.e. $P_s \neq P_t$. In addition, we are given a source model (θ, h_s) pre-trained on source data D_s, in which $\theta : \mathcal{X} \to \mathbb{R}^d$ represents a feature extractor producing d-dimensional features and $h_s : \mathbb{R}^d \to \mathcal{P}(\mathcal{Y}_s)$ is the head classifier outputting the final probability distribution of the labels, where $\mathcal{P}(\mathcal{Y}_s)$ is the space of all probability distributions over $\mathcal{Y}_s$.

Our transferability definition is based on a popular form of neural network based transfer learning (illustrated in Figure 2), known as *Retrain head*. It keeps the weights of source feature extractor θ frozen and retrains a new head classifier h_t for target task [12, 35, 42, 25]. The *ground-truth* of transferability can be represented by the empirical transfer performance on the target task, i.e., retrain the source model on target data and then evaluate the classification accuracy on its testing set. We can define the empirical transferability as follows.

Definition 1 *The empirical transferability from source task S to target task T is measured by the expected accuracy of the retrained (θ, h_t) on the testing set of target task:*

$$\mathrm{Trf}(S \to T) = \mathbb{E}\left[\mathrm{acc}(y_t, x_t; \theta, h_t)\right], \tag{1}$$

which indicates how well the source model θ performs on target task T. [38]

Although empirical transferability can be the golden standard of describing how easy it is to transfer knowledge from a source task to a target task, it is computationally expensive to obtain. Therefore, *analytical transferability* aims to effectively approximate empirical transferability, without relying on training a new network.

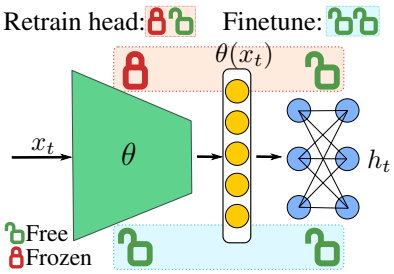

Figure 2. Illustration of two neural network based transfer learning methods, i.e., Retrain head and Finetune.

It is worth mentioning that another type of transfer learning is referred as *Finetune* [2, 14], i.e., update the feature extractor θ and the new head classifier h_t simultaneously. Compared to *Retrain head*, *Finetune* trade-offs transfer efficiency for better target accuracy and it requires more target data to avoid overfitting. As in previous analytical studies [24, 38, 5], in this paper, we pay more attention to the *Retrain head* method by working directly in the feature space determined by the source feature extractor θ. Thus the performance of current analytical transferability metrics on the finetuned model (θ_t, h_t) are generally worse than that of *Retrain head* for the same tasks. Nevertheless, experiments under the *Finetune* setting (Section 4.4) show that our OTCE score outperforms previous metrics [24, 38, 5].

3.2. Preliminary of Optimal Transport

Optimal Transport (OT) theory originated from the Monge problem in 1781, and then the Kantorovich relaxation [18] was proposed to make the Optimal Transport theory a powerful approach to leverage the underlying space for comparing distributions, shapes and point clouds [29]. The OT problem considers a complete and separable metric space $\mathcal{X}$, along with continuous or discrete probability measures $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ [3]. The Kantorovich relaxation of OT problem is defined as:

$$OT(\alpha, \beta) \triangleq \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} c(x, z) d\pi(x, z), \tag{2}$$

where $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a cost function, and $\Pi(\alpha, \beta)$ is a set of couplings, i.e., joint probability distributions over the space $\mathcal{X} \times \mathcal{X}$ with marginal distributions α, β, that is,

$$\Pi(\alpha, \beta) \triangleq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) | P_{1\#}\pi = \alpha, P_{2\#}\pi = \beta\}. \tag{3}$$

When the $c(x, z) = d_{\mathcal{X}}(x, z)^p$ of some $p \geq 1$, $W_p(\alpha, \beta) \triangleq OT(\alpha, \beta)^{1/p}$ is denoted as the p-Wasserstein distance.

In practice, we rarely know the true marginal distributions α, β. Instead, we usually compute the discrete empirical distributions $\hat{\alpha} = \sum_{i=1}^m \mathbf{a}_i \delta_{\mathbf{x}^i}, \hat{\beta} = \sum_{i=1}^n \mathbf{b}_i \delta_{\mathbf{z}^i}$, where

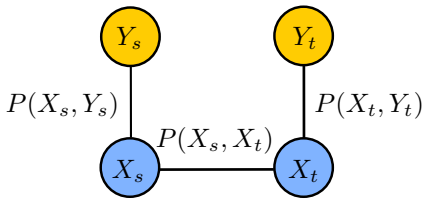

Figure 3. The probability graph model of the source and target task.

$\mathbf{a}, \mathbf{b}$ are vectors in the probability simplex. And the cost function in Equation (2) can simply be represented as an $m \times n$ cost matrix $\mathbf{C}$, where $\mathbf{C}_{ij} = c(\mathbf{x}^i, \mathbf{z}^j)$.

Furthermore, OT can be efficiently solved via the Sinkhorn algorithm [11] by adding an entropic regularizer to the objective function in Equation (2). The entropic regularized OT has been used for domain adaptation to compute the optimal mapping of input from the source domain to the target domain [10]. Alvarez *et al*. [3] also adopts OT to geometrically evaluate the distance between datasets.

3.3. OTCE Score

The motivation of our OTCE (Optimal Transport based Conditional Entropy) score is decomposing the overall difference between two classification tasks into *domain difference* and *task difference*. To this end, we adopt OT to evaluate the domain difference for its advantages in computing directly from finite empirical samples and capturing the underlying geometry of data. More importantly, by solving the OT problem between source and target data, we can obtain an optimal coupling matrix of samples, revealing the pair-wise optimal matching under a given distance metric.

From a probabilistic point of view, the coupling matrix is a non-parametric estimation of the joint probability of the source and target latent features $P(X_s, X_t)$. We model the relationship between the source and the target data according to the following simple Markov random field: $Y_s - X_s - X_t - Y_t$ (shown in Figure 3), where label random variables Y_s and Y_t are only dependent on X_s and X_t, respectively, i.e, $P(Y_s, Y_t|X_s, X_t) = P(Y_s|X_s)P(Y_t|X_t)$. Furthermore, we can derive the empirical joint probability distribution of source and target label sets,

$$P(Y_s, Y_t) = \mathbb{E}_{X_s, X_t}[P(Y_s|X_s)P(Y_t|X_t)]. \tag{4}$$

We consider this joint probability distribution to some extent represents the task difference, since the goodness of class-to-class matching may intuitively reveal the hardness of transfer. Inspired by Tran *et al*. [38] who use Conditional Entropy (CE) $H(Y_t|Y_s)$ to describe class-to-class matching quality over the same input instances, we consider it as a reasonable metric to evaluate task difference once we learn the soft correspondence between source and target features $P(X_s, X_t)$ via optimal transport. Finally, we define our analytical transferability metric OTCE as a weighted combination of domain difference and task difference.

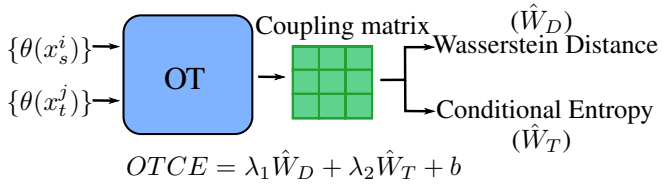

Figure 4. Illustration of our proposed OTCE transferability metric.

Figure 4 shows the overview of our proposed transferability metric. The computation process of OTCE score is described in following steps.

Step1: Compute domain difference. In our problem, we adopt the OT definition with entropic regularization [11] to facilitate the computation:

$$OT(D_s, D_t) \triangleq \min_{\pi \in \Pi(D_s, D_t)} \sum_{i,j=1}^{m,n} c(\theta(x_s^i), \theta(x_t^j))\pi_{ij} + \epsilon H(\pi), \tag{5}$$

where $c(\cdot, \cdot) = \| \cdot - \cdot \|_2^2$ is the cost metric, and π is the coupling matrix of size $m \times n$, and $H(\pi) = -\sum_{i=1}^{m}\sum_{j=1}^{n} \pi_{ij} \log \pi_{ij}$ is the entropic regularizer with $\epsilon = 0.1$. The OT problem above can be solved efficiently by Sinkhorn algorithm [11] to produce an optimal coupling matrix π^*. Thus the *domain difference* W_D can be represented by the commonly used 1-Wasserstein distance, denoted as:

$$W_D = \sum_{i,j=1}^{m,n} \|\theta(x_s^i) - \theta(x_t^j)\|_2^2 \pi_{ij}^*. \tag{6}$$

Step2: Compute task difference. Based on the optimal coupling matrix π^*, we can compute the empirical joint probability distribution of source and target label sets, and the marginal probability distribution of source label set, denoted as:

$$\hat{P}(y_s, y_t) = \sum_{i,j: y_s^i = y_s, y_t^j = y_t} \pi_{ij}^*, \tag{7}$$

$$\hat{P}(y_s) = \sum_{y_t \in \mathcal{Y}_t} \hat{P}(y_s, y_t). \tag{8}$$

Note that Equation (7) is the empirical estimation of Equation (4) for all pairs of source and target samples. Then we can compute the Conditional Entropy (CE) to represent *task difference* W_T,

$$\begin{aligned} W_T &= H(Y_t|Y_s) = H(Y_s, Y_t) - H(Y_s) \\ &= -\sum_{y_t \in \mathcal{Y}_t} \sum_{y_s \in \mathcal{Y}_s} \hat{P}(y_s, y_t) \log \frac{\hat{P}(y_s, y_t)}{\hat{P}(y_s)}. \end{aligned} \tag{9}$$

Here we explain why CE can be used to measure task difference. As the testing accuracy of the target task can be well indicated by the training log-likelihood score $l_T(\theta, h_t)$

if the model is not overfitted, we can define an alternative empirical transferability metric to Equation (1) as follows:

$$\begin{aligned}\widetilde{\mathrm{Trf}}(S \to T) &= l_T(\theta, h_t) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log P(y_t^i | x_t^i; \theta, h_t).\end{aligned} \tag{10}$$

So the following relationship is obtained,

$$\widetilde{\mathrm{Trf}}(S \to T) \geq l_S(\theta, h_s) - H(Y_t|Y_s). \tag{11}$$

Proof is detailed in [38]. $l_S(\theta, h_s)$ is a constant after the source model training, so the lower bound of transferability is determined by the Conditional Entropy (CE). In other words, the empirical transferability can be attributed to CE.

However, in cross-domain setting, CE alone is not sufficient to estimate empirical transferability as discussed in the Supplementary Section 3. One reason could be that there exists inherent uncertainty in estimating the joint distribution of source and target features through empirical samples, so we need to capture such uncertainty through domain difference. Thus the following step is to combine *domain difference* and *task difference* to obtain our OTCE score.

Step3: Compute OTCE score. Intuitively, we model the OTCE score as a linear combination of *domain difference* and *task difference*:

$$OTCE = \lambda_1 \hat{W}_D + \lambda_2 \hat{W}_T + b, \tag{12}$$

where λ_1, λ_2 are weighting coefficients for standardized *domain difference* $\hat{W}_D$ and *task difference* $\hat{W}_T$ respectively, and b is the bias term. Choosing the optimal weights is a challenging task since the importance of $\hat{W}_D$ and $\hat{W}_T$ may be different for various cross-domain configurations, as described in Section 4.7.

Consequently, we learn the coefficients for current specified source and target domains utilizing several auxiliary tasks. More specifically, we sample several pairs of source and target tasks, and compute their *domain differences, task differences* and *empirical transferability* as the transfer experience. Least square fitting is used to obtain the adjusted λ_1, λ_2, b. While the OTCE score can be generalized to higher order polynomial, we favor linear model since it is fast to compute (with analytical solution) and more interpretable. After obtaining the fitted model, we can use the OTCE score to predict transfer accuracy for any source-target task pair in the current cross-domain setting.

4. Experiments

We first evaluate our OTCE score on the largest-to-date cross-domain dataset, DomainNet [28], and another popular dataset Office31 [34] by computing the Pearson correlation coefficient between OTCE score and the empirical transferability (transfer accuracy). Three different transfer settings are considered, namely the *standard setting*, the *fixed category set size setting*, and the *few-shot setting*. Then we investigate the applications of our transferability metric in source model selection and multi-source feature fusion. Finally, we study the effects of the number of auxiliary task on the performance of our transferability metric. More discussions are included in the Supplementary.

4.1. Datasets

We generate collections of classification tasks by sampling different sets of categories from two existing cross-domain image datasets:

DomainNet [28] contains six domains (styles) of images, i.e., *Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R) and Sketch (S)*, each covering 345 common object categories. We exclude the *Infograph* domain due to its noisy annotations. It is worth mentioning that categories in DomainNet are severely imbalanced, i.e., the number of images per category ranges from 8 to 586. To eliminate the influence of imbalanced data in obtaining the empirical transferability, we limit the number of instances per category to be at most ≤ 100 in all target tasks.

Office31 [34] is a common benchmark dataset for domain adaptation algorithms on three domains: *Amazon (A), DSLR (D)* and *Webcam (W)*. It contains 4,110 images of 31 categories of objects typically found in an office environment.

4.2. Evaluation on Standard Setting

We define a standard cross-domain cross-task setting to evaluate the correlation (measured by Pearson correlation coefficient like [24, 38]) between our proposed OTCE score and the transfer accuracy. We compare performances with recent analytical transferability metrics LEEP [24], NCE [38] and H-score [5]. As the original NCE assumes that the source and target tasks are different labels on the same instances, we follow the modified implementation by [24], i.e., use the source model to predict the dummy source label for target data.

To generate source tasks, we obtain a 44-category and a 15-category classification tasks for DomainNet and Office31 respectively through random sampling. Then we train 8 source models (5 for DomainNet, 3 for Office31) for different domains on the defined source tasks initialized using an ImageNet-pretrained [33] ResNet-18 [15] model.

For target tasks, we randomly sample 100 classification tasks from each target domain. The number of categories range from 10-100 for DomainNet, and 10-31 for Office31, respectively. In each transfer configuration, we select one domain as the source domain, and consider others as target domains. Thus in this setting, we totally conduct $5 \times 4 \times 100 = 2000$ cross-domain cross-task transfer tests on DomainNet, and $3 \times 2 \times 100 = 600$ tests on Of-

Table 1. Quantitative comparisons evaluated by Pearson correlation coefficients between transferability metrics and transfer accuracy under cross-domain cross-task transfer settings, including standard setting (Section 4.2), fixed category set size setting (Section 4.3) and few-shot setting (Section 4.4). Superscript * denotes $p > 0.001$.

Transferring type	Dataset	Experimental setting			OTCE	LEEP[24]	NCE[38]	H-score[5]
		Source domain	Target domain	Data property				
Retrain head	DomainNet	C	P,Q,R,S	standard	**0.969**	0.919	0.787	0.864
		P	C,Q,R,S	standard	**0.968**	0.886	0.812	0.858
		Q	C,P,R,S	standard	**0.963**	0.942	0.935	0.843
		R	C,P,Q,S	standard	**0.972**	0.892	0.851	0.870
		S	C,P,Q,R	standard	**0.960**	0.952	0.954	0.882
	Office31	A	D,W	standard	**0.829**	0.805	0.796	0.590
		D	A,W	standard	**0.880**	0.857	0.849	0.441
		W	A,D	standard	**0.863**	0.811	0.804	0.489
				average	**0.926**	0.883	0.849	0.730
Retrain head	DomainNet	C	P,Q,R,S	fixed category set size	**0.757**	0.614	0.535	-0.599
		P	C,Q,R,S	fixed category set size	**0.712**	0.480	0.418	-0.541
		Q	C,P,R,S	fixed category set size	**0.352**	0.213	0.269	-0.288
		R	C,P,Q,S	fixed category set size	**0.639**	0.465	0.440	-0.100^*
		S	C,P,Q,R	fixed category set size	**0.435**	0.381	0.427	-0.302
				average	**0.579**	0.431	0.418	-0.346
Retrain head	DomainNet	C	P,Q,R,S	few-shot	**0.920**	0.843	0.713	0.767
		P	C,Q,R,S	few-shot	**0.924**	0.812	0.737	0.807
		Q	C,P,R,S	few-shot	**0.852**	0.836	0.825	0.786
		R	C,P,Q,S	few-shot	**0.937**	0.787	0.744	0.814
		S	C,P,Q,R	few-shot	**0.922**	0.886	0.884	0.834
	Office31	A	D,W	few-shot	**0.840**	0.803	0.793	0.640
		D	A,W	few-shot	**0.933**	0.923	0.930	0.413
		W	A,D	few-shot	**0.927**	0.920	0.926	0.277^*
Finetune	DomainNet	C	P,Q,R,S	few-shot	**0.699**	0.333	0.153^*	0.406
		P	C,Q,R,S	few-shot	**0.766**	0.414	0.309	0.554
		Q	C,P,R,S	few-shot	**0.663**	0.623	0.635	0.607
		R	C,P,Q,S	few-shot	**0.854**	0.288	0.226	0.511
		S	C,P,Q,R	few-shot	**0.681**	0.514	0.526	0.481
	Office31	A	D,W	few-shot	**0.319**	0.210^*	0.204^*	0.173^*
		D	A,W	few-shot	**0.939**	0.865	0.896	0.186^*
		W	A,D	few-shot	**0.947**	0.875	0.883	-0.002^*
				average	**0.820**	0.670	0.627	0.476

fice31. To determine the coefficients λ_1, λ_2, b of OTCE, we randomly select 10% target tasks as the auxiliary for each cross-domain configuration (specified by the set of source and target domains involved), and others are used for testing. The empirical transferability of each target task is the testing accuracy after training the source model on target data with SGD optimizer and cross entropy loss for 100 epochs. Table 1 (upper part) shows the numerical results of comparing our proposed OTCE score with LEEP, NCE and H-score. Figure 5 (the first row) visualizes the correlations of transferability metrics and empirical transferability (ground truth) on the test data. Both Table 1 and Figure 5 clearly demonstrate that OTCE score achieves higher correlation with the ground truth across all domain configurations, with about 5%, 9% and 27% gain compared to LEEP, NCE and H-score respectively.

4.3. Evaluation on Fixed Category Set Size

Analyzing the experimental results of Section 4.2, we find that transfer accuracy drops with the increasing of category set size (number of categories), shown in the Supplementary Section 5. A larger category set generally makes it more difficult to learn the target task well under the same training setting for a given source model. 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. To show OTCE score indeed captures these subtle variations, we design a more challenging experiment where all target category set sizes are the same. Specifically, we sample 100 target tasks with $category\_set\_size = 50$ for each target domain, and follow the training strategy described in the standard setting (Section 4.2). Results shown in Table 1 (middle part) and Figure 5 (the third row) demonstrate that our proposed OTCE score

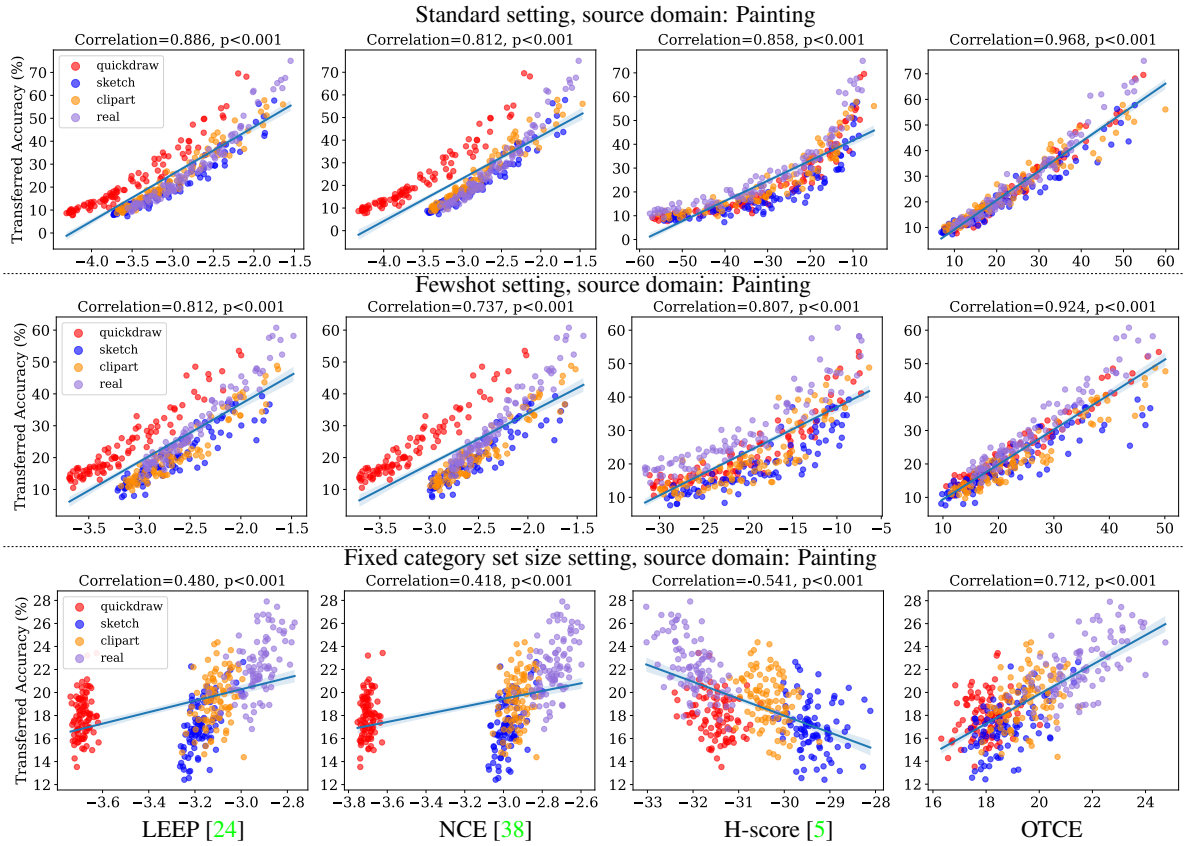

Figure 5. Visualization of correlations between empirical transferability (transfer accuracy through Retrain head) and analytical transferability metrics, including LEEP, NCE, H-score and our proposed OTCE score. Each row shows the correlations under standard setting, few-shot setting and fixed category set size setting respectively, where source domain is *Painting* and target domains are *Clipart, Quickdraw, Real, Sketch*. Points in the figure represent different target tasks. It can be seen that our OTCE score shows significantly better correlations with empirical transferability.

outperforms other transferability metrics by a large margin, with an average 34% and 39% correlation gain compared to LEEP and NCE respectively. We also note that H-score failed in this difficult setting, since correlation coefficients are negative where they should be positive.

4.4. Evaluation on Few-shot Setting

The few-shot learning problem [37] is a common application scenario of transfer learning, since training from scratch using only few-shot samples (e.g., 10 samples per category) can easily overfit, while transferring representations from a highly related source model can greatly improve the generalization of the target task. Thus it is necessary to test our transferability metric under the few-shot setting. Specifically, few-shot setting is different from the standard setting (Section 4.2) in two aspects. On one hand, we limit each category only containing 10 samples. On the other hand, we also study the correlation of transferability metrics and the transfer accuracy obtained through *Finetune*. It is worth mentioning that all the aforementioned analytical transferability metrics rely on the feature representation inferred by the source feature extractor. Therefore, it is more challenging to require transferability metrics are still highly correlated with the finetuned accuracy. Despite these restrictions, results shown in Table 1 (lower part) and Figure 5 (the second row) demonstrate that OTCE score is consistently better than LEEP, NCE and H-score with 22%, 39% and 83% correlation gain respectively.

4.5. Application for Source Model Selection

Selecting the best pretrained source model for a target task from a given set of source models is one of the most common applications of transferability metrics. In this experiment, we adopt 100 target tasks for a specified target domain as in Section 4.2. And for each target task, there are four source models pretrained on other domains. We want to evaluate whether the source model showing highest transferability score has the highest transfer accuracy on target task. If so, we consider that transferability score successfully predicts the best source model. Finally, we calculate the ratio of successful predictions. We compare the prediction accuracy among OTCE, LEEP and NCE. H-score is

Table 2. Quantitative comparisons of source model selection accuracy (%) among transferability metrics on DomainNet.

Method	Target domain					average
	C	P	Q	R	S	
LEEP[24]	31.1	26.7	5.6	97.8	**100.0**	52.2
NCE[38]	**41.1**	**94.4**	2.2	**100.0**	**100.0**	67.5
OTCE	**41.1**	93.3	**97.8**	**100.0**	**100.0**	**86.4**

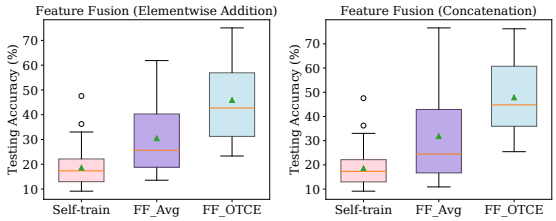

Figure 6. Testing accuracy comparisons among 'Self-train' (directly training on target data), 'FF_Avg' (average fusion) and 'FF_OTCE' (fusion weighted by OTCE score).

omitted since it does not produce any meaningful result in this experiment. Quantitative comparisons shown in Table 2 show that our OTCE score achieves top results in predicting the best source model.

4.6. Application for Multi-Source Feature Fusion

We test OTCE score on a multi-source feature fusion problem, which is another application scenario of transferability, i.e., one can transfer multiple source models to a target task by merging their inferred features together to obtain a fused representation [16]. A simple but effective fusion approach is element-wise addition or concatenation of source features. A new head classifier can be trained by taking the fused representation as input to produce the final output. However, different source models may result in different transfer performance on the target task. Thus simple average fusion is unable to effectively exploit the most useful information provided by source models. Consequently, we apply the OTCE score to weight the feature fusion for better transfer performance.

In this experiment, we sample 50 target tasks in *Real* domain of DomainNet dataset from the few-shot setting (Section 4.4). Then we employ 4 source models trained on other domains respectively to perform feature fusion targeting to these target tasks. We use a softmax function to normalize the OTCE scores of four source models to obtain the fusion coefficients in range $[0, 1]$, and then multiply source features respectively. We consider two methods of merging features, i.e., element-wise addition and concatenation. Results shown in Figure 6 demonstrate that feature fusion weighted by our OTCE score achieves the highest testing accuracy on target tasks as expected. Heuristically, our proposed transferability metric OTCE score can be an effective tool for multi-source transfer learning.

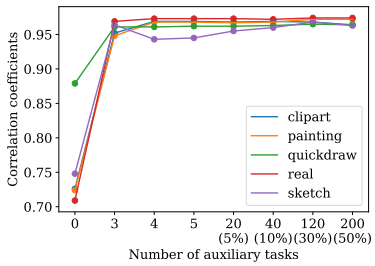

Figure 7. Study of how different number of auxiliary tasks affects the final correlations between OTCE score and empirical transferability. Polygonal lines represent different source domains from DomainNet as described in Section 4.2.

4.7. Number of Auxiliary Tasks

Auxiliary tasks are used to determine the coefficients λ_1, λ_2 and b in Equation (12) through least square fitting. We analyze the effect of auxiliary tasks on OTCE correlation using DomainNet. As shown in Figure 7, we plot the correlation between OTCE score and empirical transferability against the number of auxiliary tasks among all target tasks in each transfer setting. Note that the first data point $number = 0$ (i.e. no auxiliary training) represents the correlation using the pre-defined coefficients $\lambda_1 = \lambda_2 = -0.5$. This experiment demonstrates that learning the coefficients with auxiliary tasks for different cross-domain setting is necessary to maintain the robustness of OTCE score. Nevertheless, we can still achieve high correlation performance using only few auxiliary tasks. Moreover, we further discuss only using domain difference or task difference to characterize transferability and analyze the learned coefficients in the Supplementary Section 4.

5. Conclusion

In this study, we investigated the analytical transferability estimation problem under the general setting of cross-domain cross-task transfer learning. Our proposed transferability metric, OTCE score, characterizes the transferability between source and target tasks based on their *domain difference* and *task difference*, which can be explicitly evaluated in a unified framework. Extensive experiments demonstrate that OTCE score is more robust than other existing analytical transferability methods for capturing the uncertainty in the actual transfer performance under the cross-domain cross-task setting. For applications, we also showed through simple case studies that the OTCE score is a suitable metric to select the best source model in transfer learning and to determine feature weights in multi-source feature fusion for multi-task learning. In future works, we will explore more applications of OTCE score, such as utilizing the domain difference and task difference to support the training procedure in cross-domain cross-task transfer learning problems, e.g. open-set domain adaptation.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Int. Conf. Comput. Vis.*, pages 6430–6439, 2019. 1, 2

[2] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision*, pages 329–344. Springer, 2014. 3

[3] David Alvarez-Melis and Nicolò Fusi. Geometric dataset distances via optimal transport. *arXiv preprint arXiv:2002.02923*, 2020. 3, 4

[4] Haitham Bou Ammar, Eric Eaton, Matthew E Taylor, Decebal Constantin Mocanu, Kurt Driessens, Gerhard Weiss, and Karl Tuyls. An automated measure of mdp similarity for transfer in reinforcement learning. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014. 1

[5] 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. 1, 2, 3, 5, 6, 7

[6] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 1, 2

[7] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006. 2

[8] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003. 1, 2

[9] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008. 1, 2

[10] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016. 4

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 4

[12] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014. 3

[13] Eric Eaton, Terran Lane, et al. Modeling transfer relationships between learning tasks for improved inductive transfer. pages 317–332, 2008. 1

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Saihui Hou, Xu Liu, and Zilei Wang. Dualnet: Learn complementary features for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 502–510, 2017. 8

[17] Brendan Jou and Shih-Fu Chang. Deep cross residual learning for multitask visual recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 998–1007, 2016. 3

[18] LV Kantorovich. On the translocation of masses, cr (dokl.) acad. *Sci. URSS (NS)*, 37:199, 1942. 2, 3

[19] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017. 3

[20] Ying Lu, Lingkun Luo, Di Huang, Yunhong Wang, and Liming Chen. Knowledge transfer in vision recognition: A survey. *ACM Computing Surveys (CSUR)*, 53(2):1–35, 2020. 2

[21] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009. 1, 2

[22] Andreas Maurer. Transfer bounds for linear feature learning. *Machine learning*, 75(3):327–350, 2009. 2

[23] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016. 3

[24] 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. 1, 2, 3, 5, 6, 7, 8

[25] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 3

[26] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. 2

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 1, 2

[28] 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, 2, 5

[29] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2, 3

[30] Lorien Y Pratt. Discriminability-based transfer between neural networks. In *Advances in neural information processing systems*, pages 204–211, 1993. 1

[31] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2017. 3

[32] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 3

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[34] 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. 2, 5

[35] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 3

[36] Jivko Sinapov, Sanmit Narvekar, Matteo Leonetti, and Peter Stone. Learning inter-task transferability in the absence of target task samples. pages 725–733, 2015. 1

[37] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2019. 1, 7

[38] 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. 1, 2, 3, 4, 5, 6, 7, 8

[39] Wei Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. Transfer learning via learning to transfer. In *International Conference on Machine Learning*, pages 5085–5094, 2018. 2

[40] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 1, 2

[41] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3712–3722, 2018. 1, 2

[42] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3