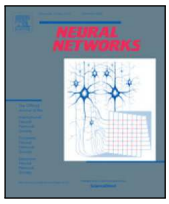
Neural Networks 181 (2025) 106796

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Wasserstein task embedding for measuring task similarities

Xinran Liu [a,*], Yikun Bai [a], Yuzhe Lu [b], Andrea Soltoggio [c], Soheil Kolouri [a]

[a] Computer Science Department, Vanderbilt University, 2201 W End Ave, Nashville, 37235, TN, United States
[b] School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, 15213, PA, United States
[c] School of Computer Science, Loughborough University, Epinal Way, Loughborough, LE11 3TU, UK

ARTICLE INFO

Keywords:
Task embedding
Dataset similarity
Optimal transport
Continual learning
Transfer learning

ABSTRACT

Measuring similarities between different tasks is critical in a broad spectrum of machine learning problems, including transfer, multi-task, continual, and meta-learning. Most current approaches to measuring task similarities are architecture-dependent: (1) relying on pre-trained models, or (2) training networks on tasks and using forward transfer as a proxy for task similarity. In this paper, we leverage the optimal transport theory and define a novel task embedding for supervised classification that is model-agnostic, training-free, and capable of handling (partially) disjoint label sets. In short, given a dataset with ground-truth labels, we perform a label embedding through multi-dimensional scaling and concatenate dataset samples with their corresponding label embeddings. Then, we define the distance between two datasets as the 2-Wasserstein distance between their updated samples. Lastly, we leverage the 2-Wasserstein embedding framework to embed tasks into a vector space in which the Euclidean distance between the embedded points approximates the proposed 2-Wasserstein distance between tasks. We show that the proposed embedding leads to a significantly faster comparison of tasks compared to related approaches like the Optimal Transport Dataset Distance (OTDD). Furthermore, we demonstrate the effectiveness of our embedding through various numerical experiments and show statistically significant correlations between our proposed distance and the forward and backward transfer among tasks on a wide variety of image recognition datasets.

1. Introduction

Learning from a broad spectrum of tasks and transferring knowledge between them is a cornerstone of intelligence, and primates perfectly exemplify this characteristic. Modern Machine Learning (ML) is rapidly moving toward multi-task learning, and there is great interest in methods that can integrate, rapidly adapt, and seamlessly transfer knowledge between tasks. When learning from multiple possibly heterogeneous tasks, it is essential to understand the relationships between the tasks and their fundamental properties. It is, therefore, highly desirable to define (dis)similarity measures between tasks that will allow one to cluster tasks, have better control over the forward and backward transfer, and ultimately require less supervision for learning tasks.

There has been an increasing interest in assessing task similarities and their relationship with forward and backward knowledge transfer among tasks. For instance, various recent works look into the selection of good source tasks/models for a given target task to maximize the forward transfer to the target task (Achille et al., 2019; Bao et al., 2019; Bhattacharjee et al., 2020; Fifty et al., 2021; Zamir et al., 2018). Others have demonstrated the relationship between negative backward transfer (i.e., catastrophic forgetting) and task similarities (Nguyen et al., 2019).

Many existing methods for measuring task similarities depend on the choice of model(s), architecture(s), and the training process (Achille et al., 2019; Gao & Chaudhari, 2021; Khodak, Balcan, & Talwalkar, 2019; Leite & Brazdil, 2005; Nguyen, Hassner, Seeger, & Archambeau, 2020; Venkitaraman, Hansson, & Wahlberg, 2020; Zamir et al., 2018). For example, Gao and Chaudhari (2021), Venkitaraman et al. (2020), Zamir et al. (2018) use pre-trained task specified models to measure a notion of forward transfer and define it as task similarity. Achille et al. (2019) embed tasks into a vector space that relies on a partially-trained network. Khodak et al. (2019) use the optimal parameters as a proxy for each task and Leite and Brazdil (2005) use the learning curves of a pre-specified model to measure task similarities. Besides being model-dependent, these approaches are often computationally expensive as they involve training deep models (or require pre-trained models).

Model-agnostic task similarity measures provide a fundamentally different approach to quantifying task relationships (Alvarez-Melis & Fusi, 2020; Ben-David, Blitzer, Crammer, & Pereira, 2006). These methods often measure the similarity between tasks as a function of the

* Corresponding author.
E-mail address: xinran.liu@vanderbilt.edu (X. Liu).

https://doi.org/10.1016/j.neunet.2024.106796
Received 4 January 2024; Received in revised form 13 September 2024; Accepted 7 October 2024
Available online 15 October 2024

similarity between the joint or conditional input/output distributions, sometimes also taking the loss function into account. The classic theoretical results for such similarity measures (Batu, Fortnow, Rubinfeld, Smith, & White, 2000; Ben-David et al., 2006) focus on information theoretic divergences between the source and target distributions. More recently, Optimal Transport (OT) based approaches (Alvarez-Melis & Fusi, 2020; Tan, Li, & Huang, 2021; Xu, Yang, Liu, Zhang, & Liu, 2022) have shown promise in modeling task similarities. Notably, Alvarez-Melis and Fusi (2020) approach measuring task similarities through the lens of a hierarchical OT (Yurochkin, Claici, Chien, Mirzazadeh, & Solomon, 2019) where they solve an inner OT problem to calculate the label distance between the class-conditional distributions of two supervised learning tasks. The label distance is then incorporated into the transportation cost of an outer OT problem, resulting in a distance between two datasets that integrates both sample and label discrepancies. Tan et al. (2021) treats the optimal transport plan between the input distributions of two tasks as a joint probability distribution and use conditional entropy to measure the difference between the two tasks. One major shortcoming of these OT-based approaches is their computational complexity. These methods require the pairwise calculation of OT (or entropy regularized OT) between different tasks, which can be prohibitively expensive in applications requiring frequent evaluations of task similarities, e.g., in continual learning.

We propose a novel OT-based task embedding for supervised learning problems that is model-agnostic and computationally efficient. On the one hand, our proposed approach is similar to Achille et al. (2019) and Peng, Li, and Saenko (2020), which embed datasets into a vector space in which one can easily measure the difference between tasks, e.g., via the Euclidean distance between embedded vectors. On the other hand, our approach is inspired by the Optimal Transport Dataset Distance (OTDD) (Alvarez-Melis & Fusi, 2020) framework, and it essentially provides a Euclidean embedding for a hierarchical OT-based distance between tasks. To calculate such a task embedding, we use the Wasserstein embedding framework (Kolouri, Naderializadeh, Rohde, & Hoffmann, 2020; Wang, Slepčev, Basu, Ozolek, & Rohde, 2013). Importantly, our approach alleviates the need for pairwise calculation of OT problems between tasks, turning it into a more desirable solution than previously proposed methods.

Contributions. We propose a computationally efficient and model-agnostic task embedding, denoted as Wasserstein Task Embedding (WTE), in which the Euclidean distance between embedded vectors approximates a hierarchical OT distance between the tasks. We provide extensive numerical experiments and demonstrate that: (1) our calculated distances between embedded tasks are highly correlated with the OTDD distance (Alvarez-Melis & Fusi, 2020), (2) our proposed embedding and similarity calculation is significantly faster than the OTDD distance, and (3) our proposed similarity measure provides strong and statistically significant correlation with both forward and backward transfer.

2. Related work

Model-based task similarity. Most existing approaches to measuring task similarity are model-dependent and use forward transferability as a proxy for similarity. Zamir et al. (2018) use pre-trained models on source tasks and measure their performance on a target task to obtain an asymmetric notion of similarity between source and target tasks. Following Zamir et al. (2018)'s work, Dwivedi and Roig (2019) measure the transferability in a more efficient manner by applying the Representation Similarity Analysis (RSA) between the trained models (e.g. DNNs) from different tasks. Similarly, Nguyen et al. (2020) assume the source and target tasks share the same set of inputs but have different sets of labels, and estimate the transferability by the empirical conditional distribution of target labels given the inputs computed by a pre-trained model on the source task.

Another class of approaches embed the tasks into a vector space and then define the (dis)similarity on the embedded vector representations. Achille et al. (2019) discuss processing data (images) through a partially trained "probe network" and obtain vector embedding by computing the Fisher information matrix (FIM). The (dis)similarity of two tasks is then computed from the difference between the FIMs. Similarly, Peng et al. (2020) propose a domain (labeled dataset) to vector technique. In particular, given a domain, they feed the data to a pre-trained CNN to compute the Gram matrices of the activations of the hidden convolutional layers, and apply feature disentanglement to extract the domain-specified features. Concatenation of the diagonal entries of Gram matrices and the domain-specified features gives the final domain embedding. These methods, however, highly rely on the pre-trained models and training process, and lack theoretical guarantees. On the opposite side of the spectrum is directly measuring the discrepancy between domains.

Discrepancy measures of domains. Over the years, numerous notions of discrepancy to measure the (dis)similarity of datasets (domains) were proposed, including L_1-norm (Batu et al., 2000), generalized Kolmogorov–Smirnov distance (Devroye, Györfi, & Lugosi, 1996), and loss-oriented discrepancy distance (Mansour, Mohri, & Rostamizadeh, 2009). In the context of domain adaptation, generalized Kolmogorov–Smirnov distance (later known as the $\mathcal{A}$-distance) is a principled notation of discrepancy, which is a relaxation of total variation. Another widely used distance is the Maximum Mean Discrepancy (MMD) (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2006), which captures the (dis)similarity of the embedding of distribution measures in a reproducing kernel Hilbert space. Pan, Tsang, Kwok, and Yang (2010) propose to learn transfer components across domains in reproducing kernel Hilbert space using MMD, and show that the subspace spanned by these transfer components preserves data properties. Such domain discrepancy methods, however, cannot take labels into account, and thus may not be enough to reflect the similarity of tasks.

Optimal transport based task similarity. In recent years, metrics rooted in the optimal transport problem, e.g., the "Wasserstein distance" (Villani, 2009, 2021) have attracted growing interest in the machine learning community. Wasserstein distance is a rigorous metric of probability measures endowed with desired statistical convergence behavior, in contrast to other classical discrepancies (e.g. KL-divergence, total variation, JS-divergence, Hellinger distance, Maximum mean discrepancy, etc.).

Alvarez-Melis and Fusi (2020) propose a notion of distance between two datasets in a supervised learning setting. They introduce Optimal Transport Dataset Distance (OTDD) based on the OT theory, which can be thought as a hierarchical OT distance where the transportation cost measures the distance between samples as well as labels. With the assumption that the label-induced distributions can be approximated by Gaussians, the distance between labels is defined as the Bures-Wasserstein distance.

Tan et al. (2021) introduce another OT-based method to measure the transferability, named OTCE (Optimal Transport Conditional Entropy) score. In particular, they first use the entropic optimal transport to estimate domain differences and then use the optimal coupling between the source and target distributions to compute the conditional entropy of the target task given source task. The OTCE is defined by the linear combination of the OT distance and the conditional entropy, whose coefficients are fitted by auxiliary tasks. To overcome this major limitation of dependency on auxiliary tasks, Tan, Zhang, Li, Huang, and Zhang (2024) propose a faster and auxiliary-free variant of OTCE, named F-OTCE, which estimates transferability by first solving an OT problem between source and target distributions, and then just using the optimal coupling to compute the Negative Conditional Entropy between source and target labels.

Both OTDD and OTCE (including F-OTCE) were shown to be effectively aligned with forward transfer, however, the computation of

Algorithm 1 Multidimensional Scaling

Input: $\mathcal{X} = \{x_n\}_{n=1}^N$, $D = [d(x_i, x_j)]_{i,j}$, l

1: $B = -\frac{1}{2}(id_{N\times N} - \frac{1}{N}\mathbb{1}_{N\times N})D(id_{N\times N} - \frac{1}{N}\mathbb{1}_{N\times N})$, where $id_{N\times N}$ is the $N \times N$ identity matrix, $\mathbb{1}_{N\times N}$ is $N \times N$ matrix of all ones.
2: Eigen-decomposition $B = V\Lambda V^T$
3: Rearrange Λ into $\hat{\Lambda}$ with descending order of variances
4: Rearrange V into $\hat{V}$ in correspondence with $\hat{\Lambda}$
5: $\hat{\Lambda}_{(l)} = \hat{\Lambda}[:l, :l]$; $\hat{V}_{(l)} = \hat{V}[:, :l]$
6: **return** $\psi(\mathcal{X}) = \hat{\Lambda}_{(l)}^{\frac{1}{2}}\hat{V}_{(l)}$

the pairwise Wasserstein distances/optimal coupling among increasing number of datasets remains expensive. This hinders the application of these methods to problems where one needs to perform nearest dataset retrieval frequently (e.g., memory replay approaches in continual learning).

Computation Cost of OT Distance. Calculating the Wasserstein distance involves solving an n^2 dimension linear programming and the computational cost is $\mathcal{O}(n^3\log(n))$ for a pair of n-size empirical distributions. To facilitate the computation, one common method is adding entropic regularization (Cuturi, 2013; Peyré, Cuturi, others, 2017), by which the original linear programming problem is converted into a strictly convex problem. By applying the Sinkhorn-Knopp algorithm (Chizat, Peyré, Schmitzer, & Vialard, 2018; Peyré, Cuturi, others, 2017) to find an ϵ- accurate solution, the computational complexity reduces to $\mathcal{O}(n^2\log(n)/\epsilon^3)$ (Altschuler, Niles-Weed, & Rigollet, 2017). However, this technique suffers a stability-accuracy trade-off. When the regularity coefficient is high, the objective is biased toward the entropy term; when it is small, the Sinkhorn algorithm will not be numerically stable.

3. Preliminaries

3.1. Multidimensional scaling (MDS)

Multidimensional scaling (MDS) (Cox & Cox, 2008) is a non-linear dimensionality reduction approach that embeds N samples into an l-dimensional Euclidean space while preserving their pairwise distances. Given a set of high-dimensional data $\mathcal{X} = \{x_n\}_{n=1}^N$ and the proximity matrix $D \in \mathbb{R}^{N\times N}$, where $D_{i,j} = d(x_i, x_j)$, and $d(\cdot,\cdot)$ denotes the metric in $\mathcal{X}$, the goal of MDS is to construct a distance-preserving map from $\mathcal{X}$ to a lower-dimensional Euclidean space $\mathbb{R}^l$. Depending on the objective and inputs, MDS can be classified into metric MDS and non-metric MDS. Specifically, metric MDS aims to find a map $\psi : \mathcal{X} \to \mathbb{R}^l$ such that

$$\min_{\psi} \sqrt{\frac{\sum_{i,j}\left(d(x_i, x_j) - \|\psi(x_i), \psi(x_j)\|\right)^2}{\sum_{i,j} d(x_i, x_j)^2}}, \tag{1}$$

which can be solved by Algo. 1.

Note that MDS not only works for Euclidean distances, but also for other dissimilarities such as Wasserstein distances (Hamm, Henscheid, & Kang, 2022; Wang et al., 2011).

3.2. Wasserstein distances

Let μ, ν be Borel probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$ with finite pth moment, and the corresponding probability density functions are p_μ and p_ν, i.e. $d\mu = p_\mu dx$, $d\nu = p_\nu dx$. The 2-Wasserstein distance between μ and ν is defined as (Villani, 2009):

$$\mathcal{W}_2(\mu, \nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{X}} \|x - x'\|^2 d\gamma(x, x')\right)^{\frac{1}{2}}, \tag{2}$$

where $\Gamma(\mu, \nu)$ is the set of all transport plans between μ and ν, i.e. probability measures on $\mathcal{X}\times\mathcal{X}$ with marginals μ and ν. We also note that by Brenier theorem (Brenier, 1991), given two absolutely continuous probability measures μ, ν on $\mathbb{R}^d$ with densities p_μ, p_ν, there exists a convex function ϕ such that $T = \nabla\phi$ is a transport map sending μ to ν. Moreover, it is the optimal map in the Monge–Kantorovitch optimal transport problem with quadratic cost:

$$\mathcal{W}_2(\mu, \nu) = \left(\int_{\mathcal{X}} \|x - T(x)\|^2 p_\mu dx\right)^{\frac{1}{2}}, \tag{3}$$

where $T = \nabla\phi$ pushes μ to ν, denoted by $T_\#\mu = \nu$.

3.3. Wasserstein Embedding (WE)

Wasserstein Embedding (Courty, Flamary, & Ducoffe, 2017; Kolouri, Naderializadeh, Rohde, & Hoffmann, 2021; Kolouri, Tosun, Ozolek, & Rohde, 2016; Wang et al., 2013) provides a Hilbertian embedding for probability distributions such that the Euclidean distance between the embedded vectors approximates the 2-Wasserstein distance between the two distributions. Let $\{\mu_i\}_{i=0}^I$ be a set of I probability distributions over $\mathcal{X} \subseteq \mathbb{R}^d$ with densities $\{p_i\}_{i=0}^I$. We fix μ_0 as the reference measure. Assume T_i is the optimal transport map that pushes μ_0 to μ_i, the Wasserstein embedding of μ_i is through a function Φ defined as

$$\Phi(\mu_i) = (T_i - id)\sqrt{p_0} \tag{4}$$

where the id is the identity function, i.e., $id(x) = x$. Φ admits nice properties including but not limited to Kolouri et al. (2021):

1. $d(\mu_i, \mu_j) := \|\Phi(\mu_i) - \Phi(\mu_j)\|_2$ is a true metric between μ_i and μ_j, moreover, it approximates the 2-Wasserstein distance: $d(\mu_i, \mu_j) \approx \mathcal{W}_2(\mu_i, \mu_j)$.
2. In particular, $\|\Phi(\mu_i)\|_2 = \|\Phi(\mu_i) - \Phi(\mu_0)\|_2 \approx \mathcal{W}_2(\mu_i, \mu_0)$. Here we leveraged the fact $\Phi(\mu_0) = 0$.

Although these hold true for both continuous and discrete measures $\{\mu_i\}_{i=0}^I$, we focus on the (uniformly distributed) discrete setting in this paper and provide the following numerical computation details. Let $p_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\delta_{x_n^i}$, where δ_x is the Dirac delta function centered at $x \in \mathcal{X}$ and $X_i = \{x_n^i\}_{n=1}^{N_i}$ is the set of locations of non-negative mass for μ_i. Then the Kantorovich problem with quadratic cost between μ_i and μ_0 can be formulated as

$$\min_{\pi\in\Pi_i}\sum_{n=1}^{N_0}\sum_{k=1}^{N_i}\pi_{nk}\|x_n^i - x_k^0\|_2^2 \tag{5}$$

where the feasible set is

$$\Pi_i := \{\pi \in \mathbb{R}^{N_0\times N_i} | N_0\sum_{n=1}^{N_i}\pi_{nk} = N_i\sum_{k=1}^{N_0}\pi_{nk} = 1,\ \forall n, k\}. \tag{6}$$

The discrete formulation Eq. (5) corresponds to the formulation in Eq. (2) with squaring both sides. The optimal transport plan π_i^* is the minimizer of the above optimization problem, which is solved by linear program at cost $\mathcal{O}(N^3\log(N))$, N being the number of input samples. To avoid mass splitting, the barycentric projection (Wang et al., 2013) assigns each x_j^0 in the reference distribution to the center of mass it is sent to and thus outputs an approximated Monge map T_i. Then the Wasserstein Embedding for input X_i is calculated by

$$\Phi(X_i) = (T_i - X_0)/\sqrt{N_0} \in \mathbb{R}^{N_0\times d}. \tag{7}$$

One of the motivations behind Wasserstein embedding is to ameliorate the need for computing pairwise Wasserstein distances. Given M datasets, computation of $\frac{M(M-1)}{2}$ Wasserstein distances (i.e., OT problems) across all distinct pairs is impractically expensive especially when M is large, while leveraging Wasserstein embedding, it suffices to calculate only M OT problems and the pairwise Euclidean distances between the embedded distributions.

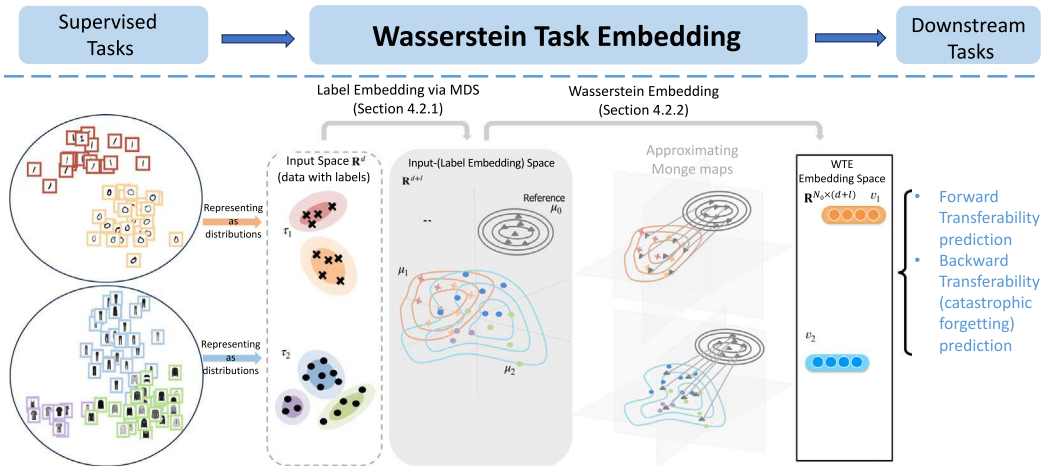


Fig. 1. Wasserstein Task Embedding framework. Given labeled task distributions τ_1 and τ_2 with input space $\mathbb{R}^d$, WTE first maps them into $\mathbb{R}^{d+l}$ as probability distributions μ_1 and μ_2 by label embedding via MDS, then apply WE to get vectors v_1 and v_2 with respect to a fixed reference measure μ_0. Here N_0 is the size of reference set.

3.4. Optimal Transport Dataset Distance (OTDD)

Let $\mathcal{X} = \{x_n \in \mathbb{R}^d\}_{n=1}^N$ be the input set with labels (classes) $\mathcal{Y} = \{y_n\}_{n=1}^N$. Following the OTDD framework (Alvarez-Melis & Fusi, 2020), let $\tau = \{(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}\}_{n=1}^N$ denote the set of data-label pairs. OTDD encodes each label y as distribution ν_y, where $\nu_y = \frac{1}{|C_y|}\sum_{x\in C_y}\delta_x$ and $C_y \subset \mathcal{X}$ is the subset of input set with the label y. The ground distance in τ is then defined by combining the Euclidean distance between the data points and the 2-Wasserstein distance between label distributions:

$$d_\tau\left((x,y),(x',y')\right) := \left(\|x-x'\|^2 + \mathcal{W}_2^2(\nu_y,\nu_{y'})\right)^{\frac{1}{2}}. \tag{8}$$

Based on this metric, the OT distance between two distributions μ_i and μ_j on τ is

$$d_{OT}(\mu_i,\mu_j) = \inf_{\pi\in\Pi(\mu_i,\mu_j)}\int_{\tau\times\tau} d_\tau(z,z')^2 d\pi(z,z'), \tag{9}$$

where $\Pi(\mu_i,\mu_j)$ denotes the set of transport plans between μ_i and μ_j. Note that Eq. (9) is a hierarchical transport problem, as the transportation cost itself depends on calculation of the Wasserstein distance. To avoid the computational cost of a hierarchical optimal transport problem, Alvarez-Melis and Fusi (2020) replace the Wasserstein distance in Eq. (8) with the Bures-Wasserstein distance (Bhatia, Jain, & Lim, 2019; Malago, Montrucchio, & Pistone, 2018), which assumes that ν_ys are Gaussian distributions. Throughout the paper, we consider only the exact-OTDD, as opposed to the entropy-regularized and other variants.

4. Method

In this section, we specify the problem setting, review the OTDD framework, and then propose our Wasserstein task embedding (WTE).

4.1. Problem setting

In supervised classification problems, tasks are represented by input-label pairs and can be denoted as $\tau \subseteq \mathcal{X}\times\mathcal{Y} = \{(x_n,y_n)\}_{n=1}^N$, where $\mathcal{X}\subseteq\mathbb{R}^d$ is the data/inputs and $\mathcal{Y}$ is the labels. We aim to define a similarity/dissimilarity measure for tasks that enable task clustering and allow for better control over the forward and backward transfer.

4.2. Wasserstein task embedding

We define a task-2-vec framework (Fig. 1) using Wasserstein embedding (WE) such that the (squared) Euclidean distance between two vectors approximates the OTDD between the original tasks, and denote this embedding by WTE. We later show in the experiment section that the Euclidean distance between the embedded task vectors is not only highly predictive of forward transferability, but also significantly correlates with the backward transferability (catastrophic forgetting).

4.2.1. Label embedding via MDS

The combination of optimal transport metric with MDS technique was first introduced as an approach to characterize and contrast the distribution of nuclear structure in different tissue classes (normal, benign, cancer, etc.) (Wang et al., 2011), and further studied in image manifold learning (Hamm et al., 2022). In short, it seeks to isometrically map probability distributions to vectors in relatively low-dimensional space. We leverage the prior work and define an approximated isometry on the label distributions by (1) calculating the pairwise Wasserstein distances and (2) applying MDS to obtain embedded vectors. We adopt the same simplification as in OTDD, that is, assuming the label distributions are Gaussians to replace Wasserstein distances with the closed form Bures-Wasserstein distance:

$$\mathcal{W}_2^2(\nu_y,\nu_{y'}) = \|u_y - u_{y'}\|_2^2 + \mathrm{Tr}(\Sigma_y + \Sigma_{y'} - 2(\Sigma_y^{\frac{1}{2}}\Sigma_{y'}\Sigma_y^{\frac{1}{2}})^{\frac{1}{2}}) \tag{10}$$

where u and Σ denote the mean and covariance matrix of Gaussian distributions. In consistence with previous notations, let us denote the label (MDS) embedding operator by ψ, then

$$\mathcal{W}_2^2(\nu_y,\nu_{y'}) \approx \|\psi(\nu_y) - \psi(\nu_{y'})\|_2^2, \tag{11}$$

where $\psi(\nu_y), \psi(\nu_{y'}) \in \mathbb{R}^l$ are vectors whose dimension l is selected to balance the trade-off between accuracy and computation cost (Tenenbaum, Silva, & Langford, 2000). Having both inputs and labels represented as vectors, we concatenate these two components and map the data-label pairs τ to $\tau' \subseteq \mathbb{R}^{d+l}$ such that

$$\begin{aligned} d_\tau\left((x,y),(x',y')\right) &\approx (\|x-x'\|_2^2 + \|\psi(\nu_y) - \psi(\nu_{y'})\|_2^2)^{\frac{1}{2}} \\ &= (\|[x,\psi(\nu_y)] - [x',\psi(\nu_{y'})]\|_2^2)^{\frac{1}{2}} \\ &= \|[x,\psi(\nu_y)] - [x',\psi(\nu_{y'})]\|_2 \end{aligned} \tag{12}$$

where $[\cdot,\cdot]$ denotes the concatenation operator and the domain $\tau' = \{[x,\psi(\nu_y)]\}_{x\in\mathcal{X}}$ is equipped with l_2 norm. Fig. 2 shows this approximation performance among labels in MNIST (LeCun & Cortes, 2005) and USPS (Hull, 1994) datasets. MDS embeddings can capture the pairwise relationships with a maximum of 7.26% error by 10-dimensional vectors.

4.2.2. Wasserstein Embedding

By Eqs. (12) and (9), OTDD can be approximated by the squared 2-Wasserstein distance between the distributions over input-(label MDS embedding) pairs, τ'. Then we leverage the Wasserstein embedding

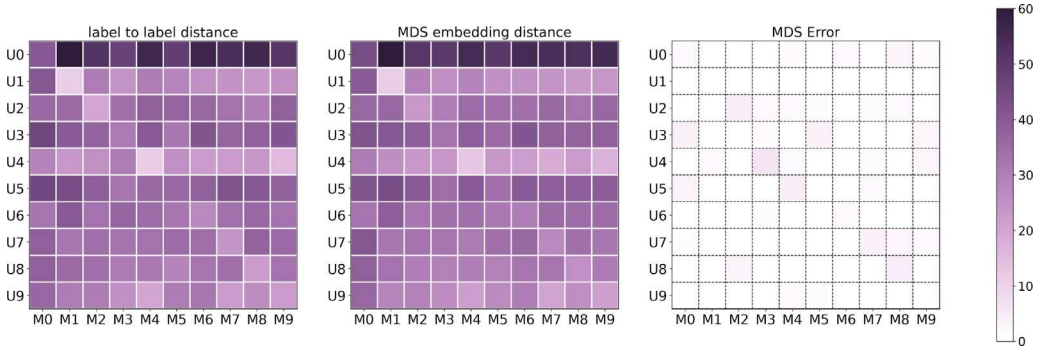


Fig. 2. Label-to-label Bures-Wasserstein distance (left) and label MDS embedding Euclidean distances (middle) between MNIST and USPS datasets, squared error is provided on the right.

Algorithm 2 Wasserstein Task Embedding

Input: $\{X_i = \{(x_n^i, y_n^i)\}_{n=1}^{N_i}\}_{i=1}^{I}$

1: Calculate label-to-label distance matrix D (Eq. (10))
2: Calculate $\psi(\nu_y) \in \mathbb{R}^l$ for all distinct labels y (Algo. 1)
3: Stack each input with its label vector: $x \to [x, \psi(\nu_y)] \in \mathbb{R}^{d+l}$
4: Fix a discrete reference distribution μ_0 on $\tau_0 = \{z_m^0\}_{m=1}^{N_0} \subset \mathbb{R}^{d+l}$ with density $p_0 = \frac{1}{N_0}\sum_{m=1}^{N_0} \delta_{z_m^0}$.
5: **for** $i \in [1, 2, \cdots, I]$ **do**
6: Calculate the optimal transport plan π^i from μ_0 to the discrete measure μ_i with distribution on $\tau_i = \{z_n^i = [x_n^i, \psi(\nu_{y_n^i})]\}_{n=1}^{N_i}$ by solving the minimization in Eq. (5).
7: Calculate the approximated Monge map $T_i = [\frac{1}{N_i}\sum_{n=1}^{N_i} \pi_{m,n}^i z_n^i]_{m=1}^{N_0}$ as a stack of vectors by barycentric projection (Wang et al., 2013) of π^i.
8: $v_i = (T_i - \tau_0)/\sqrt{N_0}$
9: **end for**
10: **return** $\{v_i\}_{i=1}^{I}$

framework to embed the updated task distributions into a Hilbert space, with the goal of reducing the cost of computing pairwise Wasserstein distances. Again, we emphasize that this can bring down the cost from quadratic to linear with the number of task distributions.

The WTE algorithm is summarized in Algo. 2. The outputs are the vector representations of input tasks with respect to a pre-determined MDS dimension and WE reference distribution.

5. Experiments

To assess the effectiveness of our WTE framework, we empirically validate the correlation between WTE distance and forward/backward transferability on several datasets. Moreover, we provide both qualitative and quantitative comparison results with OTDD, and show WTE distance is well aligned with OTDD, and meanwhile is notably faster to compute. We use the MDS toolkit in scikit-learn and the exact linear programming solver in Python Optimal Transport (POT) (Flamary et al., 2021) library for implementing WE. We carry out the distance calculations on CPU and all the model training experiments on a 24 GB NVIDIA RTX A5000 GPU.

5.1. Datasets

We conduct experiments on the following four task groups:

***NIST* task group** consists of the handwritten digits dataset MNIST (LeCun & Cortes, 2005) and its extensions EMNIST (Cohen, Afshar, Tapson, & Van Schaik, 2017), FashionMNIST (Xiao, Rasul, & Vollgraf, 2017), KMNIST (Clanuwat et al., 2018) along with USPS (Hull, 1994). We choose the MNIST split for EMNIST dataset and thus all tasks contain 10 classes of gray-scale images. All datasets have a training set of 60,000 samples and a test set of 10,000 samples, except USPS, with a total of 9298 samples. We resize the images from USPS into 28×28 pixel level to match with the others.

***Split-CIFAR100* task group** is generated by randomly splitting the CIFAR-100 (Krizhevsky, 2009) dataset with 100 image categories into 10 smaller tasks, each of which is a classification with 10 classes. There are 600 32×32 color images in the training set and 100 in the test set per class.

***Split-Tiny ImageNet* task group** follows the same splitting scheme as in Split-CIFAR100. We randomly divide the Tiny ImageNet (Le & Yang, 2015) into 10 disjoint tasks with 20 classes. Each class contains 500 training images, 50 validating images and 50 test images. For better model performance, we first rescale each sample to 256×256 and then perform a center crop to get 224×224 pixel images.

***DomainNet* task group** (Peng et al., 2019) contains 6 domains/tasks: Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R) and Sketch (S). We discard the Infograph task due to its noisy annotations. To mitigate the class imbalance, we randomly sample at most 100 images in each category.

5.2. Results

To study the transfer behaviors against the WTE distances, we fix a model architecture for each task group. Specifically, we use ResNet18 (He, Zhang, Ren, & Sun, 2016) on *NIST, Split-Tiny ImageNet and DomainNet, and ResNet34 (He et al., 2016) on Split-CIFAR100. In the forward transfer setting, for each source-target task pair, we first train the head (i.e., the classifier) of a randomly initialized backbone on the target task, and use the test performance as the baseline. Next, we adapt from a model pre-trained on the source task and finetune the head on the target task. We define the forward transferability of the source-target pair as the performance gain, i.e. error drop when adapting from the source task. To analyze backward transfer, all source tasks are trained jointly during the first phase to avoid task bias, then in the second phase the model learns only the target task and suffers from "forgetting" the previous tasks. We use the catastrophic forgetting, i.e., negative backward transfer as a measure of backward (in)transferability. In implementations of WE, the reference distribution is fixed for each task group, and is randomly generated by upsampling random images at a lower spatial resolution to entail some smooth structure.

Fig. 3 summarizes the correlation diagrams between our proposed WTE distance and the forward/backward transferability on the aforementioned three task groups. WTE distance is negatively correlated with forward transferability, and positively correlated with catastrophic forgetting. In all scenarios, the correlation is strong and statistically significant, which confirms the efficacy of WTE distance as a measure of task similarities. We also visualize the comparison between WTE distance and OTDD on the *NIST task group in Fig. 4, showing strong correlation between the two distances.

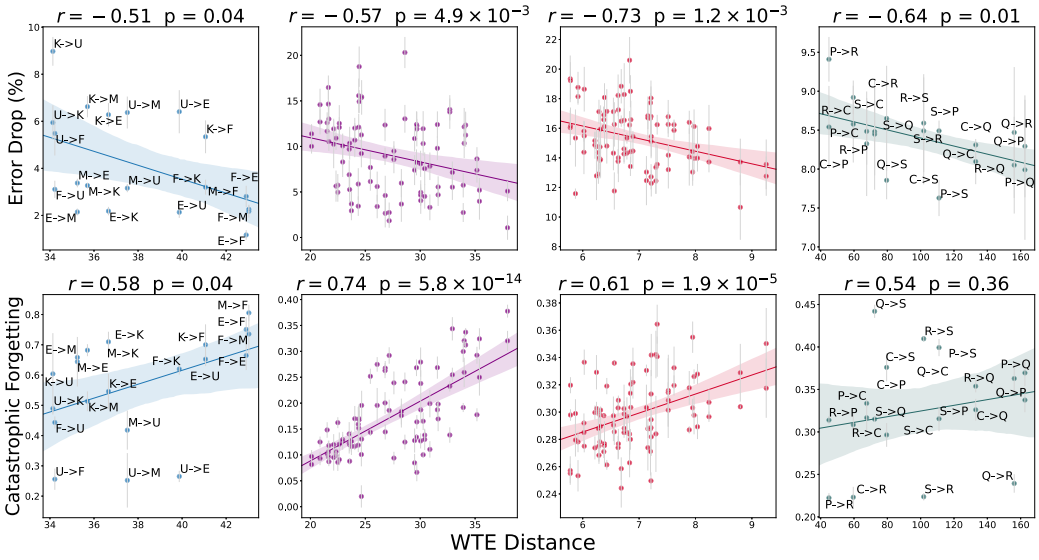


Fig. 3. (Top row) forward transfer error drop and (bottom row) catastrophic forgetting against WTE distance on *NIST, Split-CIFAR100, Split-Tiny ImageNet and DomainNet over five runs. Pearson's r and the corresponding p-value are reported on top of each experiment setting.

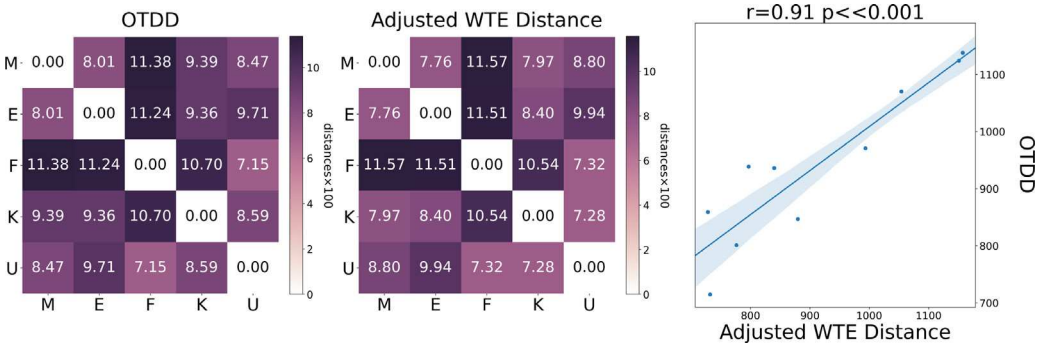


Fig. 4. Pairwise OTDD (left) and WTE distances (middle) on the *NIST task group, and their correlation diagram (right). Notice that OTDD (Eq. (9)) is the squared $\mathcal{W}_2$, we report the squared WTE distances and adjust to the same scale according to the cost function. Adjusted WTE distance is strongly correlated with OTDD, with correlation coefficient $r = 0.91$.

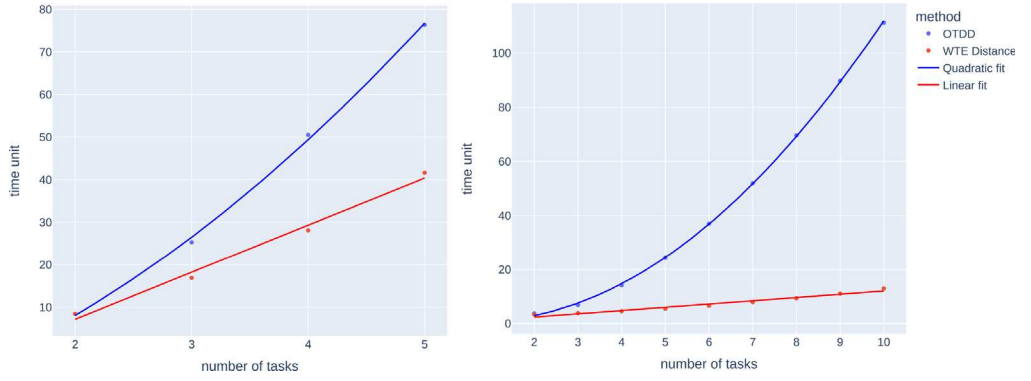


Fig. 5. Wall-clock computation time comparison on the *NIST (left) and Split-CIFAR100 (right) task groups.

Table 1
Correlation coefficients between the negative transferability with OT-based distance/score, all with p-value less than 0.05.

Task group	Source	WTE distance	OTDD	OTCE	F-OTCE
DomainNet	C	0.774	0.783	0.731	0.728
	S	0.487	0.522	0.436	0.412
	R	0.518	0.523	0.514	0.516
	Q	0.527	0.548	0.534	0.509
	P	0.722	0.721	0.713	0.711

5.3. Computation complexity

As we mentioned before, OTDD suffers from a prohibitive computational cost as the number of tasks grows large. The pairwise OTDD calculation for a set of M tasks requires $\mathcal{O}(M^2N^3\log(N))$ time in the worst case, where N is the largest number of samples among the tasks. WTE distance requires solving only M optimal transport problem, leading to $\mathcal{O}(MN^3\log(N))$ complexity. To better demonstrate the efficiency of WTE distance, we report the wall-clock time comparison on the *NIST and Split-CIFAR100 in Fig. 5.

5.4. Forward transferability correlation

We provide the Pearson correlation comparison results among the four OT-based transferability measures in Table 1 for the DomainNet task group. WTE distance approximates OTDD, and achieves comparable performance to predict forward transferability. OTCE score also shows strong correlation with forward transferability in most settings, however it is worth noting that OTCE is not completely model-agnostic as it uses model performance on auxiliary tasks to fit the linear weights. For the auxiliary task construction of OTCE in this experiment, we follow the original paper and randomly select 10% target tasks as the auxiliary for each configuration to determine the coefficients in OTCE score using least squares fitting. Besides, we take into account the effect of the number of auxiliary tasks on the OTCE performance (as illustrated in Figure 7 of Tan et al., 2021), and set the number of auxiliary tasks to be 4 to reach a good trade-off between correlation and computation cost. We have also included results for F-OTCE (Tan et al., 2024), the rapid and auxiliary-free version of OTCE. This variant demonstrates comparable performance to OTCE, with only a marginal reduction in effectiveness.

6. Conclusion

In this paper, we propose Wasserstein task embedding (WTE), a model-agnostic task embedding framework for measuring task (dis)similarities in supervised classification problems. We perform a label embedding through multi-dimensional scaling and leverage the 2-Wasserstein embedding framework to embed tasks into a vector space, in which the Euclidean distance between the embedded points approximates the 2-Wasserstein distance between tasks. We demonstrate that our proposed task embedding distance is correlated with forward and backward transfer on *NIST, Split-CIFAR100, Split-Tiny ImageNet and DomainNet task groups while being significantly faster than existing methods. In particular, we show statistically significant negative correlation between the WTE distances and the forward transfer, and positive correlation with the catastrophic forgetting (i.e. negative backward transfer). Lastly, we show the alignment of WTE distance with OTDD, but with a significant computational advantage as the number of tasks grows.

CRediT authorship contribution statement

Xinran Liu: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Conceptualization. **Yikun Bai:** Writing – original draft. **Yuzhe Lu:** Software, Investigation, Conceptualization. **Andrea Soltoggio:** Writing – review & editing, Methodology, Conceptualization. **Soheil Kolouri:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Notation list

- $\mathbb{R}^d, \mathbb{R}^l, \mathbb{R}^{d+l}, \mathbb{R}^{N_0(d+l)}$: Euclidean spaces with dimension d, l, $d+l$ and $N_0(d+l)$;
- $\mathcal{X}$: A subset in $\mathbb{R}^d$;
- $\mathcal{Y} = \{y_n\}_{n=1}^N$: Labels corresponding to input dataset $\mathcal{X} = \{x_n \in \mathbb{R}^d\}_{n=1}^N$;
- τ : The set of data-label pairs $\{(x, y) \in \mathcal{X} \times \mathcal{Y}\}$;
- $C_y \subset \mathcal{X}$: The subset in $\mathcal{X}$ with label y;
- D : Proximity matrix for MDS;
- $d(\cdot,\cdot)$: A metric on a space in the context;
- $\psi : \mathcal{X} \to \mathbb{R}^l$: Mapping by MDS;
- μ, ν : Borel probabilities on a space indicated in the context;
- p_μ, p_ν : The probability density functions corresponding to μ and ν. With slight abuse of notation we may use μ and p_μ interchangeably;
- $\mathcal{W}_2$: 2-Wasserstein metric;
- $\Gamma(\mu, \nu)$: The set of all transport plans between μ and ν, i.e. probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν;
- $id, id_{N\times N}$: Identity function and identity matrix of size $N \times N$, resp;
- $\mathbb{1}_{N\times N}$: $N \times N$ matrix of all ones;
- T : Monge map;
- $T_\# \mu$: The push-forward measure of μ by T;
- Φ : Mapping by Wasserstein embedding;
- Π : The set of discrete transport plans, in correspondence to Γ;
- π^* : Optimal transport plan;
- μ_0 : The reference measure for Wasserstein Embedding;
- N_0 : Size of μ_0, when μ_0 is discrete;
- u, Σ : Mean and covariance of Gaussian distributions;
- M : Number of supervised tasks;
- I : Number of distributions.

Data availability

Data and code are publicly available.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., et al. (2019). Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6430–6439).

Altschuler, J., Niles-Weed, J., & Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in Neural Information Processing Systems, 30*.

Alvarez-Melis, D., & Fusi, N. (2020). Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems, 33*, 21428–21439.

Bao, Y., Li, Y., Huang, S.-L., Zhang, L., Zheng, L., Zamir, A., et al. (2019). An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing* (pp. 2309–2313). IEEE.

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W. D., & White, P. (2000). Testing that distributions are close. In *Proceedings 41st annual symposium on foundations of computer science* (pp. 259–269). IEEE.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, *19*.

Bhatia, R., Jain, T., & Lim, Y. (2019). On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, *37*(2), 165–191.

Bhattacharjee, B., Kender, J. R., Hill, M., Dube, P., Huo, S., Glass, M. R., et al. (2020). P2L: Predicting transfer learning for images and semantic relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 760–761).

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, *44*, 375–417.

Chizat, L., Peyré, G., Schmitzer, B., & Vialard, F.-X. (2018). Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, *87*(314), 2563–2609.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). Deep learning for classical Japanese literature. ArXiv, arXiv:1812.01718.

Cohen, G., Afshar, S., Tapson, J., & Van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. In *2017 International joint conference on neural networks* (pp. 2921–2926). IEEE.

Courty, N., Flamary, R., & Ducoffe, M. (2017). Learning Wasserstein embeddings. arXiv preprint arXiv:1710.07457.

Cox, M. A., & Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization* (pp. 315–347). Springer.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, *26*.

Devroye, L., Györfi, L., & Lugosi, G. (1996). Parametric classification. In *A probabilistic theory of pattern recognition* (pp. 263–278). Springer.

Dwivedi, K., & Roig, G. (2019). Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12387–12396).

Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., & Finn, C. (2021). Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, *34*, 27503–27516.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., et al. (2021). POT: Python optimal transport. *Journal of Machine Learning Research*, *22*(78), 1–8, URL: http://jmlr.org/papers/v22/20-451.html.

Gao, Y., & Chaudhari, P. (2021). An information-geometric distance on the space of tasks. In *International conference on machine learning* (pp. 3553–3563). PMLR.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, *19*.

Hamm, K., Henscheid, N., & Kang, S. (2022). Wassmap: Wasserstein isometric mapping for image manifold learning. arXiv preprint arXiv:2204.06645.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hull, J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(5), 550–554. http://dx.doi.org/10.1109/34.291440.

Khodak, M., Balcan, M.-F. F., & Talwalkar, A. S. (2019). Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, *32*.

Kolouri, S., Naderializadeh, N., Rohde, G. K., & Hoffmann, H. (2020). Wasserstein embedding for graph learning. arXiv preprint arXiv:2006.09430.

Kolouri, S., Naderializadeh, N., Rohde, G. K., & Hoffmann, H. (2021). Wasserstein embedding for graph learning. In *International conference on learning representations*. URL: https://openreview.net/forum?id=AAes_3W-2z.

Kolouri, S., Tosun, A. B., Ozolek, J. A., & Rohde, G. K. (2016). A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, *51*, 453–462.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images.

Le, Y., & Yang, X. S. (2015). Tiny ImageNet visual recognition challenge.

LeCun, Y., & Cortes, C. (2005). The mnist database of handwritten digits.

Leite, R., & Brazdil, P. (2005). Predicting relative performance of classifiers from samples. In *Proceedings of the 22nd international conference on machine learning* (pp. 497–503).

Malago, L., Montrucchio, L., & Pistone, G. (2018). Wasserstein riemannian geometry of positive definite matrices. arXiv preprint arXiv:1801.09269.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. arXiv preprint arXiv:0902.3430.

Nguyen, C. V., Achille, A., Lam, M., Hassner, T., Mahadevan, V., & Soatto, S. (2019). Toward understanding catastrophic forgetting in continual learning. arXiv preprint arXiv:1908.01091.

Nguyen, C., Hassner, T., Seeger, M., & Archambeau, C. (2020). Leep: A new measure to evaluate transferability of learned representations. In *International conference on machine learning* (pp. 7294–7305). PMLR.

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, *22*(2), 199–210.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1406–1415).

Peng, X., Li, Y., & Saenko, K. (2020). Domain2vec: Domain embedding for unsupervised domain adaptation. In *European conference on computer vision* (pp. 756–774). Springer.

Peyré, G., Cuturi, M., et al. (2017). Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017–86).

Tan, Y., Li, Y., & Huang, S.-L. (2021). Qq. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15779–15788).

Tan, Y., Zhang, E., Li, Y., Huang, S.-L., & Zhang, X.-P. (2024). Transferability-guided cross-domain cross-task transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Tenenbaum, J. B., Silva, V. d., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Venkitaraman, A., Hansson, A., & Wahlberg, B. (2020). Task-similarity aware meta-learning through nonparametric kernel regression. arXiv preprint arXiv:2006.07212.

Villani, C. (2009). *Optimal transport: Old and new*: *vol. 338*, Springer.

Villani, C. (2021). *Topics in optimal transportation*: *vol. 58*, American Mathematical Soc.

Wang, W., Ozolek, J. A., Slepčev, D., Lee, A. B., Chen, C., & Rohde, G. K. (2011). An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, *30*(3), 621–631. http://dx.doi.org/10.1109/TMI.2010.2089693.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., & Rohde, G. K. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, *101*(2), 254–269.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.

Xu, R., Yang, X., Liu, B., Zhang, K., & Liu, W. (2022). Selecting task with optimal transport self-supervised learning for few-shot classification. arXiv preprint arXiv:2204.00289.

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., & Solomon, J. M. (2019). Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, *32*.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3712–3722).