Generating Synthetic Datasets by Interpolating along Generalized Geodesics

Jiaojiao Fan, and David Alvarez-Melis

Microsoft   Georgia Institute of Technology

Problem setup

Given:

The classification test dataset Q, and several training datasets $\{P_i\}$, $i = 1, 2, 3, \dots$

Question: Which dataset to choose for training purpose?

Attempts:

Use the union of $\{P_i\}$ ✗ too time-consuming, even detrimental

Train on P_i one by one ✗ catastrophic forgetting

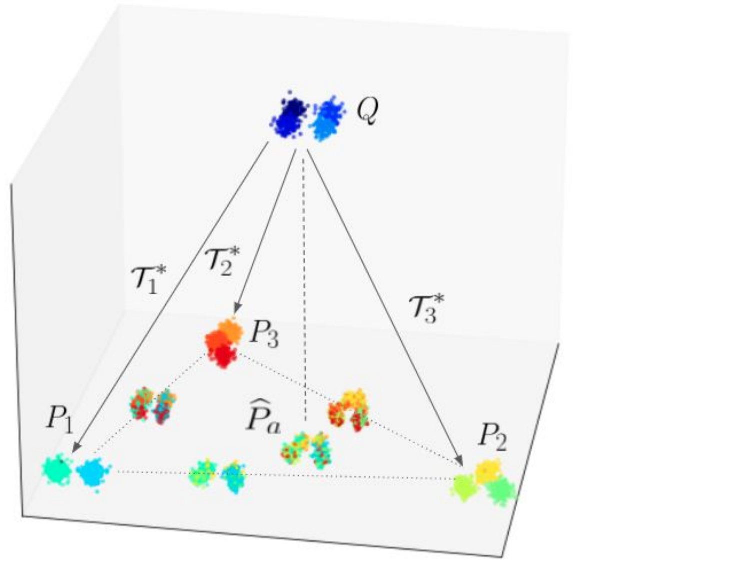
Train on a carefully chosen interpolation of $\{P_i\}$ ✓ Efficient, information loss-less

Proposed framework

Step 1: solve all the OTDD map from the reference dataset to all the training datasets

Step2: generate synthetic dataset on the generalized geodesic of all training datasets

Step 3: select the projection of test dataset as the train dataset



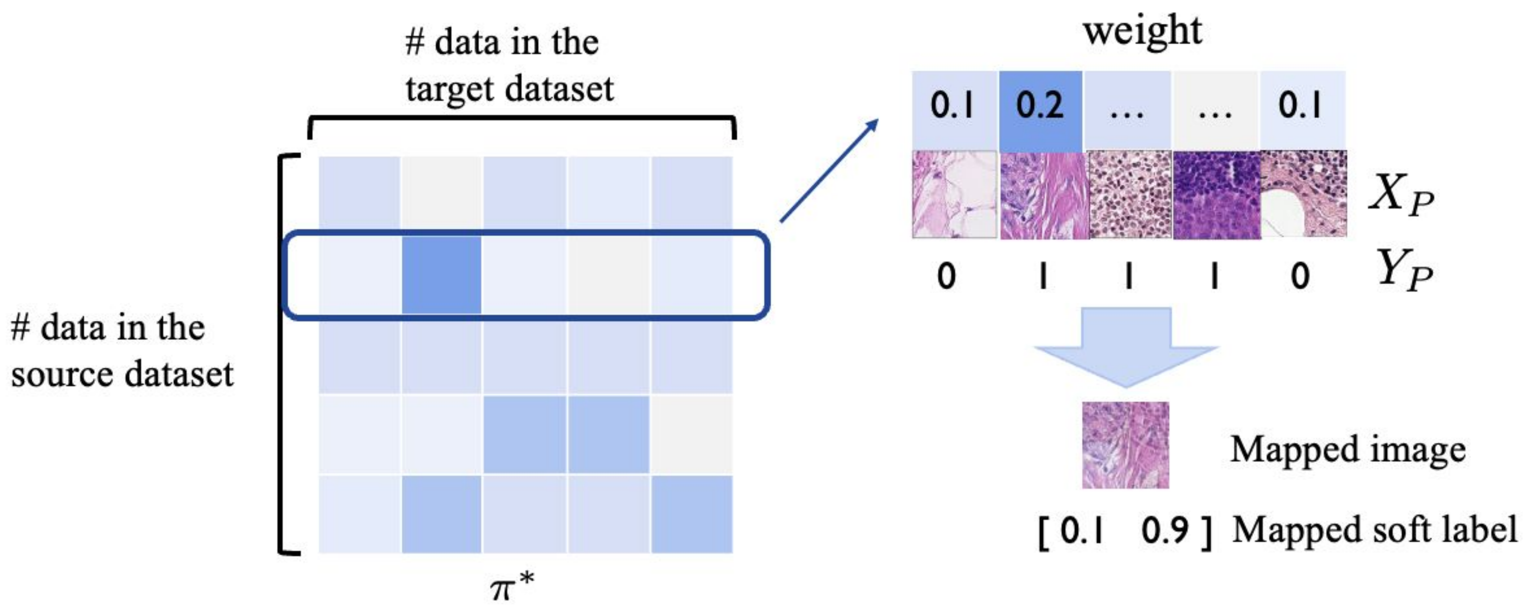
OTDD map: the optimal transport map between labeled datasets

Method 1: OTDD barycentric projection

$$\mathcal{T}_B(Z_Q) = [N_Q \pi^* X_P, N_Q \pi^* Y_P]$$

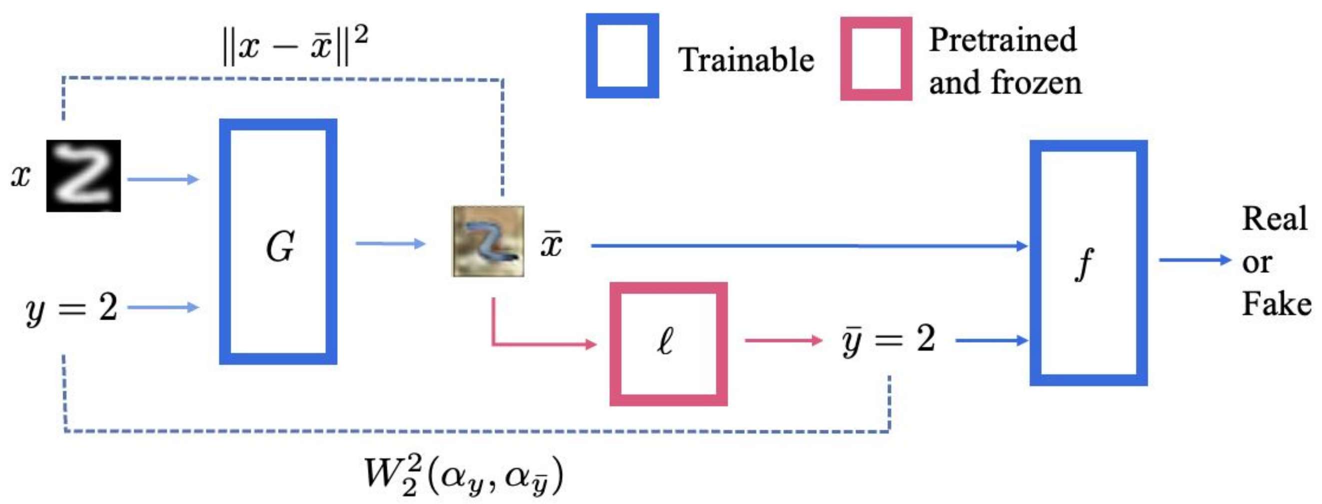
$$X_P = (x_P^{(1)}, \dots, x_P^{(N_P)})$$

$$Y_P = (y_P^{(1)}, \dots, y_P^{(N_P)})$$



Method 2: OTDD neural map

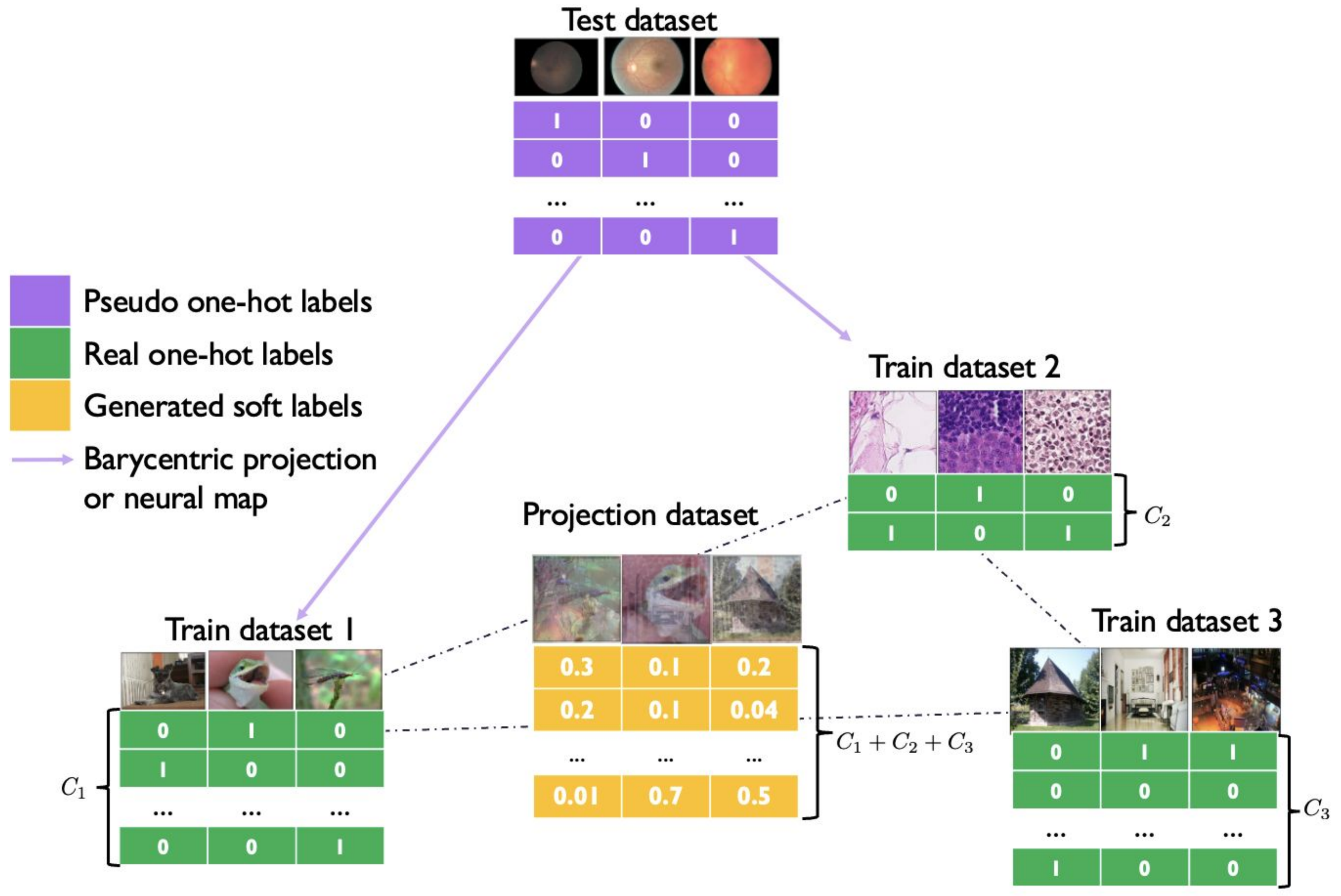
$$\mathcal{T}_N(z) = \mathcal{T}_N(x, y) = [\bar{x}; \bar{y}] = [G(z); \ell(G(z))]$$



Generalized geodesic

feature mixup: $x_a = \sum_{i=1}^{m} a_i x_i$

label mixup: $y_a = a_1 \begin{bmatrix} y_1 \\ \mathbf{0}_3 \\ \mathbf{0}_{11} \end{bmatrix} + a_2 \begin{bmatrix} \mathbf{0}_7 \\ y_2 \\ \mathbf{0}_{11} \end{bmatrix} + a_3 \begin{bmatrix} \mathbf{0}_7 \\ \mathbf{0}_3 \\ y_3 \end{bmatrix}$



The projection onto the generalized geodesic

Define approximated projection $\widehat{P}_a$ as the minimizer of function

$$\mathcal{W}^2(P_a, Q) := \sum_{i=1}^{m} a_i \mathcal{W}_{2,Q}^2(P_i, Q) - \frac{1}{2} \sum_{i \neq j} a_i a_j \mathcal{W}_{2,Q}^2(P_i, P_j),$$

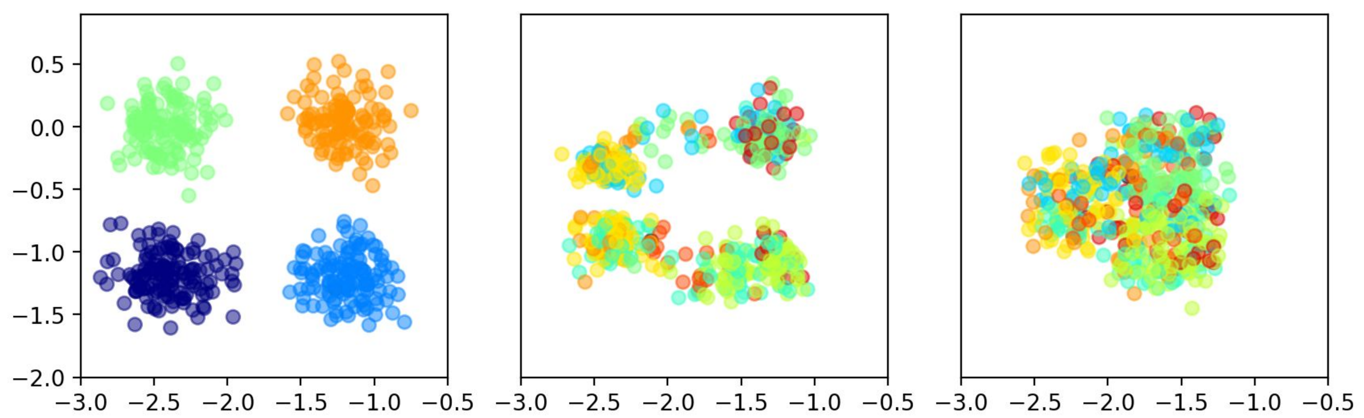
$\Rightarrow$ The minimizer is easily solvable by quadratic programming

The squared (2,Q)-dataset distance is given by

$$\mathcal{W}_{2,Q}^2(P_i, P_j) := \int \left(\|x_i - x_j\|_2^2 + W_2^2(\alpha_{y_i}, \alpha_{y_j}) \right) \mathrm{Q}(z)$$

where $[x_i; y_i] = \mathcal{T}_i^*(z)$ and $\mathcal{T}_i^*$ is the OTDD map from Q to P_i.

$\Rightarrow \mathcal{W}_{2,Q}^2$ is a valid metric



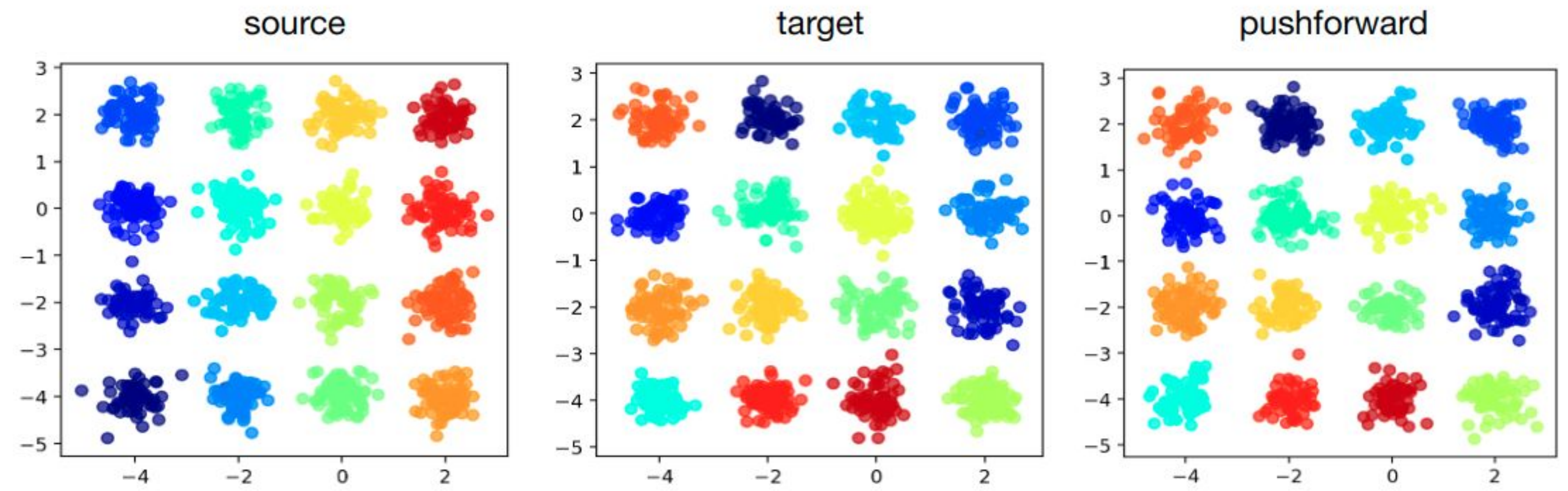
Left to right:
original dataset, projection with optimal map and random chosen map

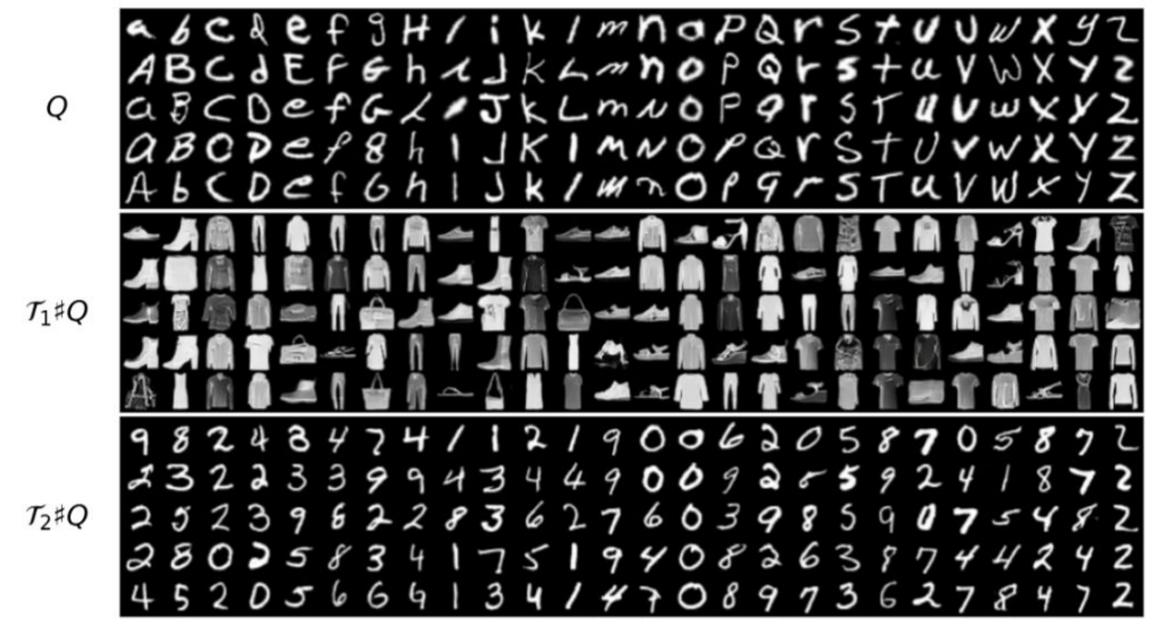
Contribution

- a novel approach to generate new synthetic classification datasets from existing ones by using geodesic interpolations, applicable even if they have disjoint label sets
- two efficient methods to compute generalize geodesics, which might be of independent interest
- empirical validation of the method in a transfer learning setting

Experiment:

Mapping between labeled datasets





Transfer learning on *NIST datasets

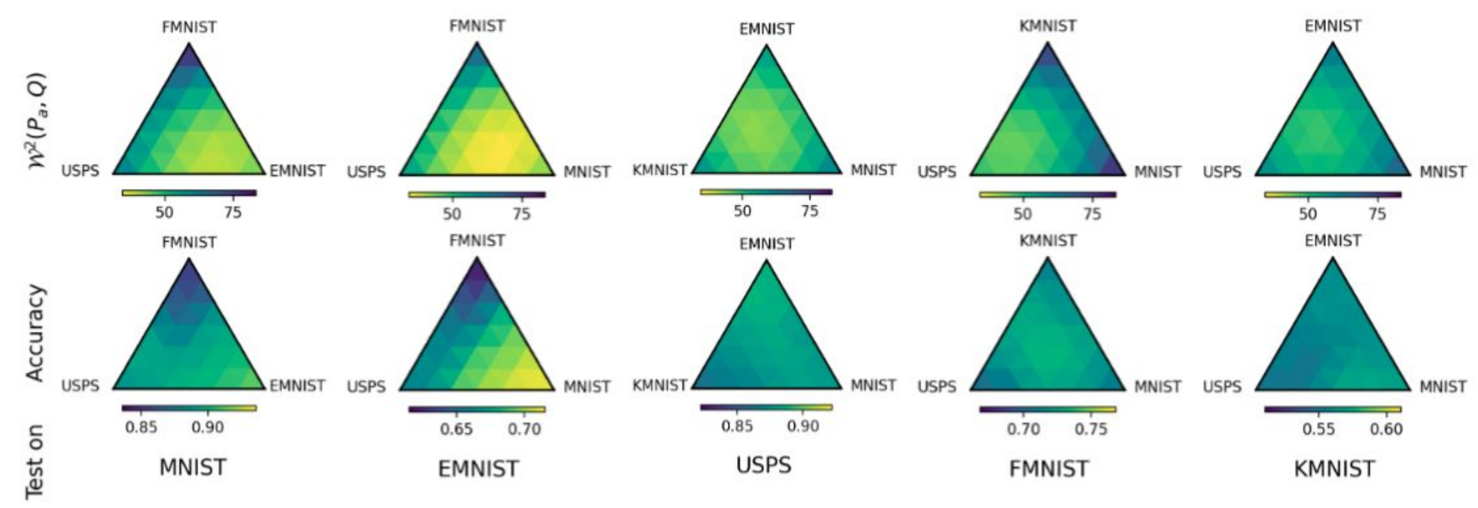


Table 1: **Pretraining on synthetic data**. Shown is 5-shot transfer accuracy (mean ± s.d. over 5 runs).

Methods	MNIST-M	MNIST	USPS	FMNIST	KMNIST	EMNIST
OTDD barycentric projection	**42.10±4.37**	**93.74±1.46**	86.01±1.50	**70.12±3.02**	**52.55±2.73**	**67.06±2.55**
OTDD neural map	40.06±4.75	88.78±3.85	83.80±1.60	70.02±2.59	50.32±3.10	65.32±1.80
Mixup	33.85±2.22	88.68±1.57	**88.61±2.00**	66.74±3.79	48.16±3.38	60.95±1.38
Train on few-shot dataset	19.10±3.57	72.80±3.10	80.73±2.07	60.50±3.07	41.67±2.11	53.60±1.18
1-NN on few-shot dataset	20.95±1.39	64.50±3.32	73.64±2.35	60.92±2.42	40.18±3.09	39.70±0.57

Transfer learning on VTAB datasets

Pre-Training	Map	Weights	Rel. Improv. (%)
CALTECH101	–	–	59.68 ± 41.44
DTD	–	–	-1.17 ± 9.52
FLOWERS102	–	–	-2.45 ± 26.25
Pooling	–	–	28.96 ± 18.29
Sub-pooling	–	–	3.00 ± 19.10
Interpolation	Mixup	uniform	33.26 ± 21.30
Interpolation	Mixup	$\hat{a}$	51.99 ± 34.10
Interpolation	OTDD	uniform	82.61 ±25.93
Interpolation	OTDD	$\hat{a}$	**95.17± 20.57**