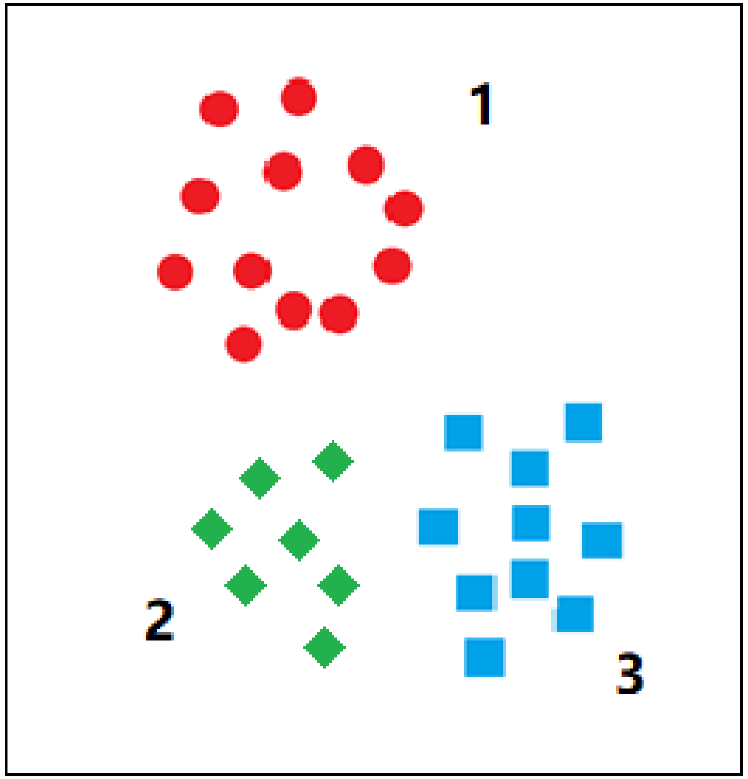
Geometric Dataset Distances via Optimal Transport

Paper Reading

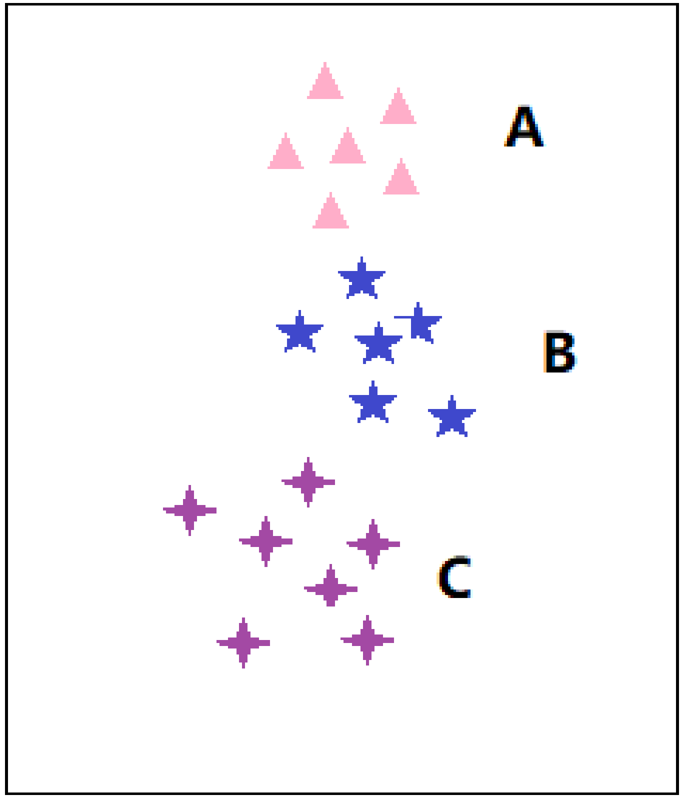
Yang Tan

2020/04/17

Dataset Distance

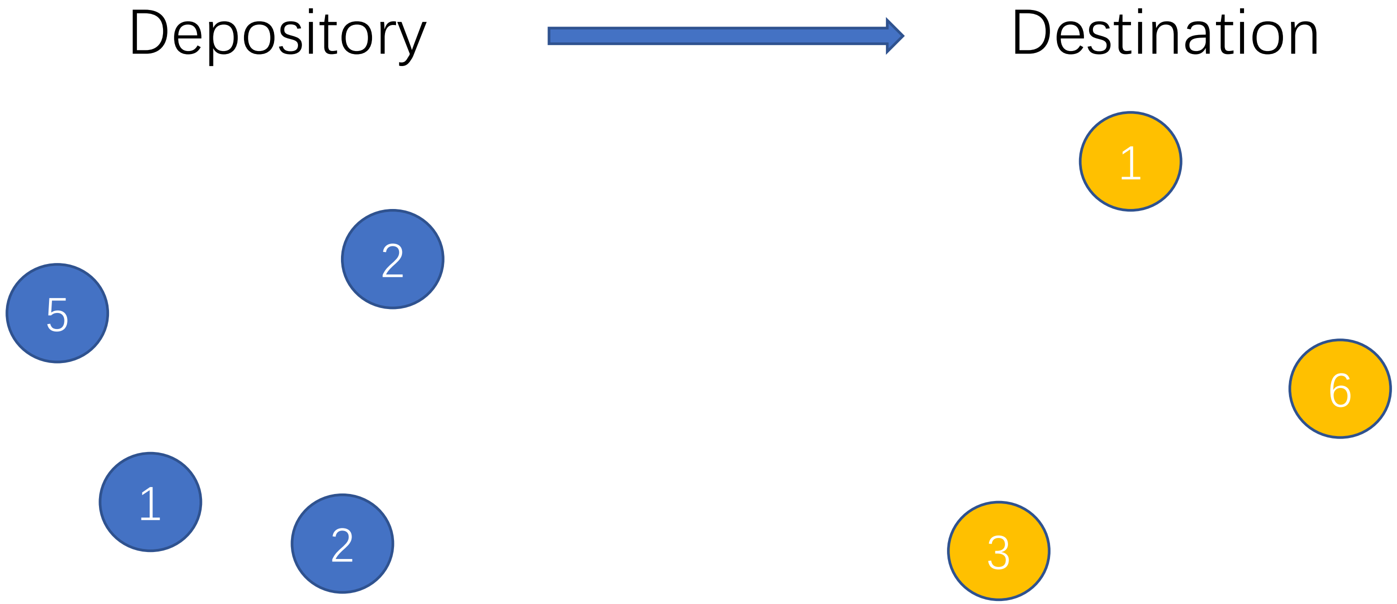


Dataset 1



Dataset 2

Optimal Transport



$$L = \arg\min_{\Gamma} \sum_{i,j=1}^{M,N} \Gamma_{ij} c\left(x_i, y_j\right)$$

Optimal Transport

- Probabilistic definition

$$L = \arg\min_{\pi} \int_x \int_y \pi(x,y)\, c(x,y) dxdy$$

- Optimal Transport Divergence

$$OT(P \| Q) = \inf_{\pi} \int_{X \times Y} \pi(x,y)\, c(x,y) dxdy$$

- K-Wasserstein distance

$$W_k(P,Q) = \inf_{\pi} \int_{X \times Y} \pi(x,y)\, \|x-y\|_k^k dxdy$$

- In this paper

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

$$c(x,y) = d_{\mathcal{X}}(x,y)^p$$

$$\mathbf{W}_p(\alpha,\beta) \triangleq \mathbf{OT}(\alpha,\beta)^{1/p}$$

Related work

- **Discrepancy**: *Ben-David et al., 2007; Mansour et al., 2009*.
- **Fisher information metric**: *Achille et al., 2019.*
- **Kolmogorov Structure Function**: *Achille et al., 2018.*
- **Optimal Transport**: *Delon & Desolneux, 2019; Dukler et al., 2019; Alvarez- Melis et al., 2018.*

Contribution

- Model agnostic
- Does not involve training
- Can compare datasets even if datasets are completely disjoint

Method

- Definitions

predictors $f : \mathcal{X} \rightarrow \mathcal{Y}$ (or conditional distributions $P(y \mid x)$), we define a dataset $\mathcal{D}$ as a set of feature-label pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ over a certain feature space $\mathcal{X}$ and label set $\mathcal{Y}$. For simplicity, we will use $z \triangleq (x, y)$ to denote these pairs, and $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ for their underlying space.

$$\mathcal{D}_A = \{(x_A^{(i)}, y_A^{(i)})\}_{i=1}^{n} \sim P_A(x, y)$$

$$\mathcal{D}_B = \{(x_B^{(j)}, y_B^{(j)})\}_{j=1}^{m} \sim P_B(x, y)$$

$$d(\mathcal{D}_A, \mathcal{D}_B) \ ?$$

Method

- Intuitively, we can define the distance as

$$d_{\mathcal{Z}}(z, z') = \left(d_{\mathcal{X}}(x, x')^p + d_{\mathcal{Y}}(y, y')^p\right)^{1/p}$$

- We can use the relationship to feature vectors to define d_y:

$$\mathcal{N}_{\mathcal{D}}(y) := \{x \in \mathcal{X} \mid (x, y) \in \mathcal{D}\}$$

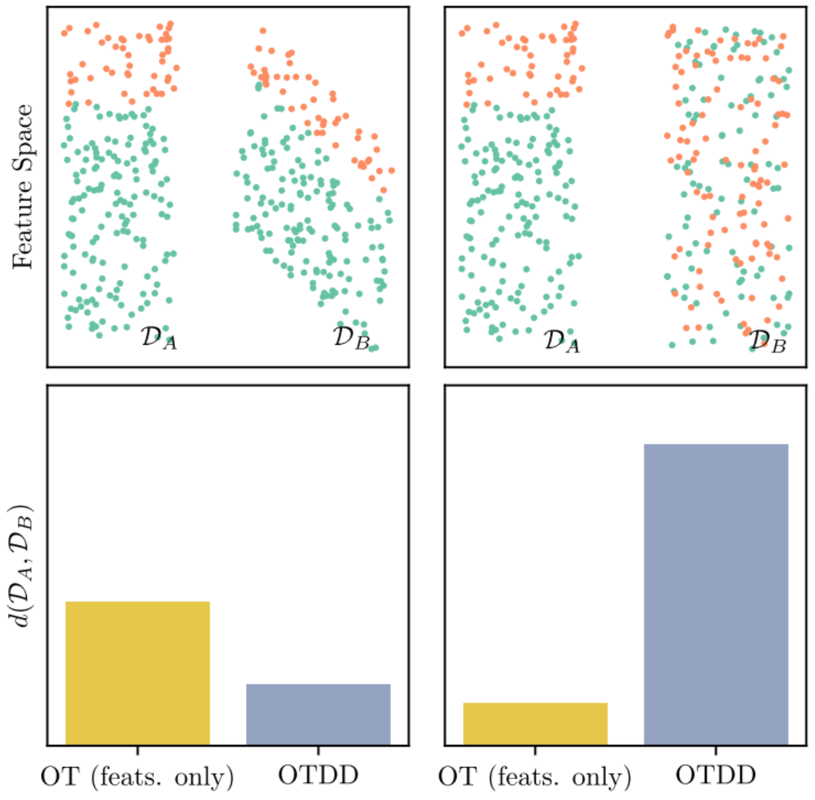
$$d(y, y') = d_{\mathcal{X}}\left(\frac{1}{n_y} \sum_{x \in \mathcal{N}_{\mathcal{D}}(y)} x, \frac{1}{n_{y'}} \sum_{x \in \mathcal{N}_{\mathcal{D}}(y')} x\right)$$

- Only measuring the mean is too simplistic for real dataset, thus consider:

$$y \mapsto \alpha_y(X) \triangleq P(X \mid Y = y)$$

$$d_{\mathcal{Z}}\left((x, y), (x', y')\right) \triangleq \left(d_{\mathcal{X}}(x, x')^p + \mathrm{W}_p^p(\alpha_y, \alpha_{y'})\right)^{\frac{1}{p}}$$

$$d_{\mathrm{OT}}(\mathcal{D}_A, \mathcal{D}_B) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{Z} \times \mathcal{Z}} d_{\mathcal{Z}}(z, z') \pi(z, z')$$



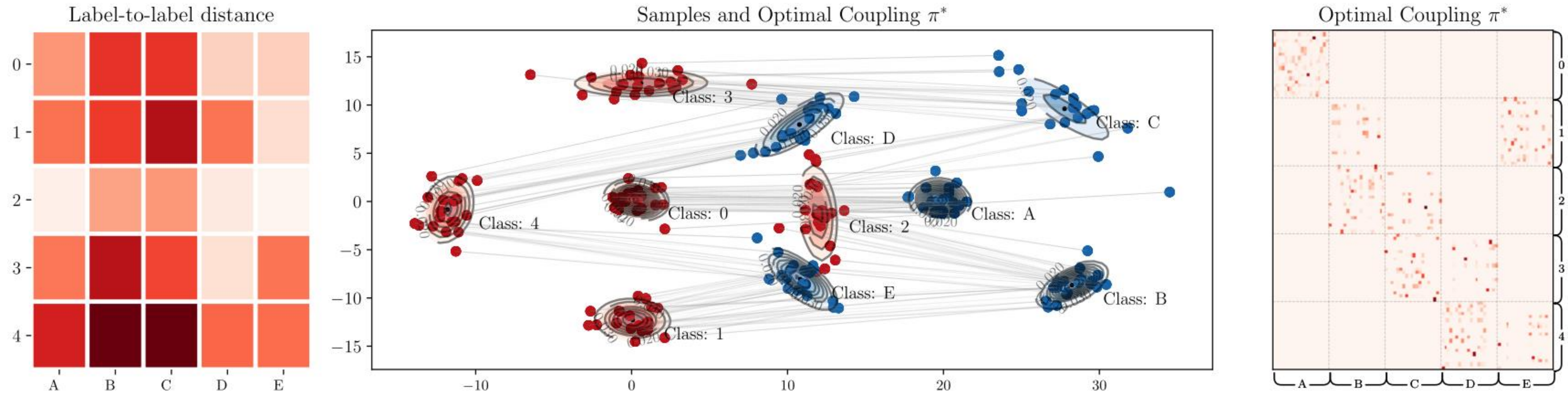
The importance of considering labels

Method

- How to describe α_y? Gaussian distribution.

$$\hat{\mu}_y \triangleq \frac{1}{n_y} \sum_{x \in \mathcal{N}_{\mathcal{D}}(y)} x; \hat{\Sigma}_y \triangleq \frac{1}{n_y} \sum_{x \in \mathcal{N}_{\mathcal{D}}(y)} (x - \hat{\mu}_y)^\top (x - \hat{\mu}_y)$$

$$\mathrm{W}_2^2(\alpha, \beta) = \|\mu_\alpha - \mu_\beta\|_2^2 + \mathrm{tr}(\Sigma_\alpha + \Sigma_\beta - 2(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}})^{\frac{1}{2}})$$



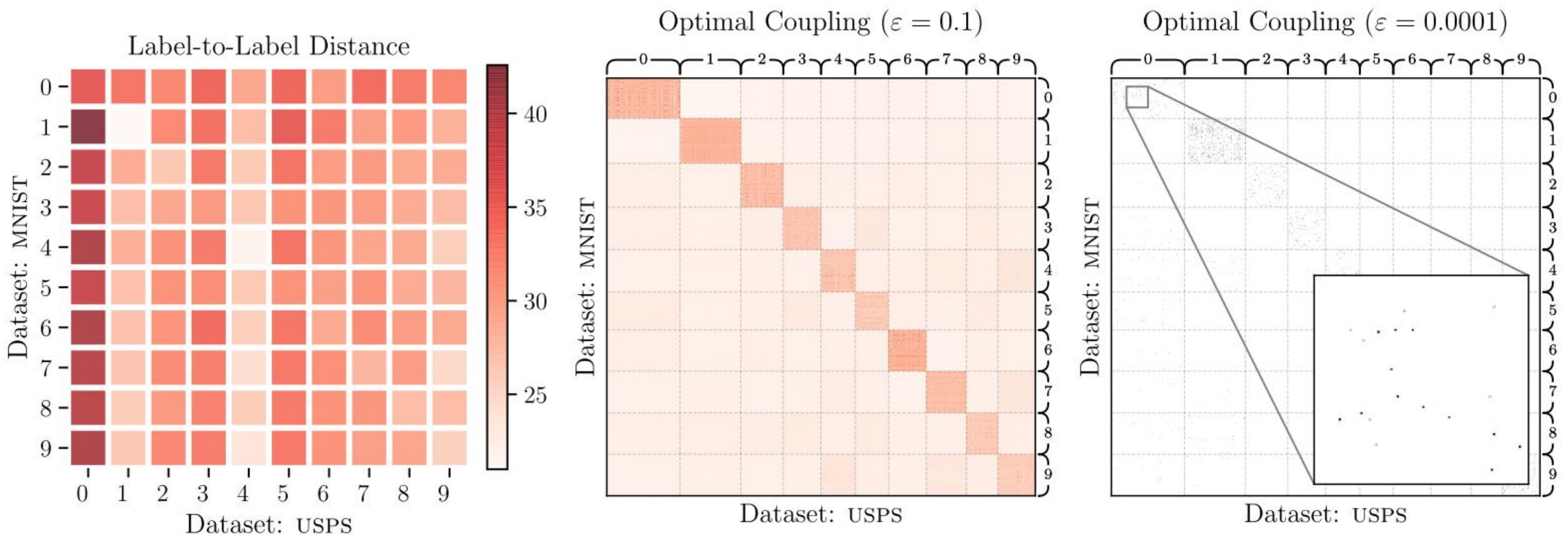
Experiments

- Datasets

Dataset	Input Dimension	Number of Classes	Train Examples	Test Examples	Source
USPS	$16 \times 16^*$	10	7291	2007	(Hull, 1994)
MNIST	28×28	10	60K	10K	(LeCun et al., 2010)
EMNIST (letters)	28×28	26	145K	10K	(Cohen et al., 2017)
KMNIST	28×28	10	60K	10K	(Clanuwat et al., 2018)
FASHION-MNIST	28×28	10	60K	10K	(Xiao et al., 2017)
TINY-IMAGENET	$64 \times 64^\ddagger$	200	100K	10K	(Deng et al., 2009)
CIFAR-10	32×32	10	50K	10K	(Krizhevsky & Hinton, 2009)
AG-NEWS	$768^\dagger$	4	120K	7.6K	(Zhang et al., 2015)
DBPEDIA	$768^\dagger$	14	560K	70K	(Zhang et al., 2015)
YELPREVIEW (Polarity)	$768^\dagger$	2	560K	38K	(Zhang et al., 2015)
YELPREVIEW (Full Scale)	$768^\dagger$	5	650K	50K	(Zhang et al., 2015)
AMAZONREVIEW (Polarity)	$768^\dagger$	2	3.6M	400K	(Zhang et al., 2015)
AMAZONREVIEW (Full Scale)	$768^\dagger$	5	3M	650K	(Zhang et al., 2015)
YAHOO ANSWERS	$768^\dagger$	10	1.4M	60K	(Zhang et al., 2015)

Experiments

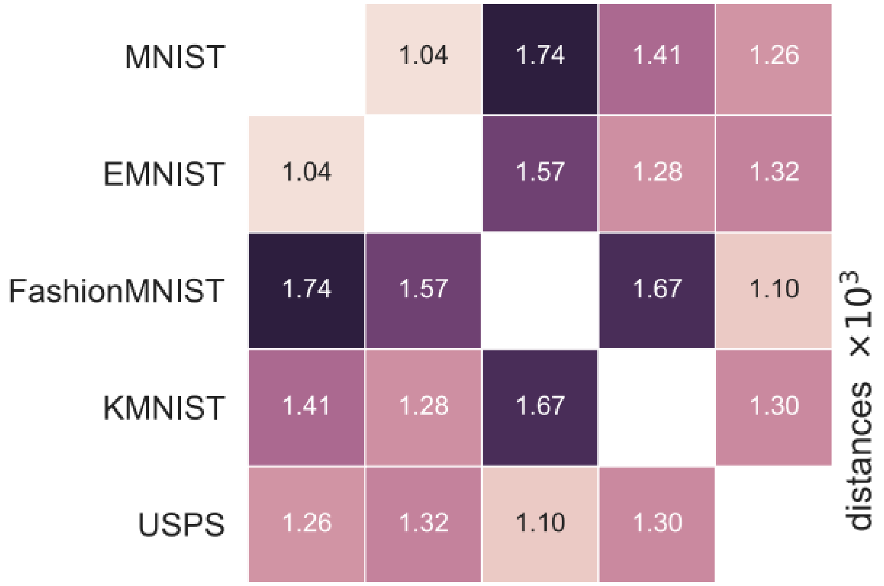
- Dataset Selection for Transfer Learning

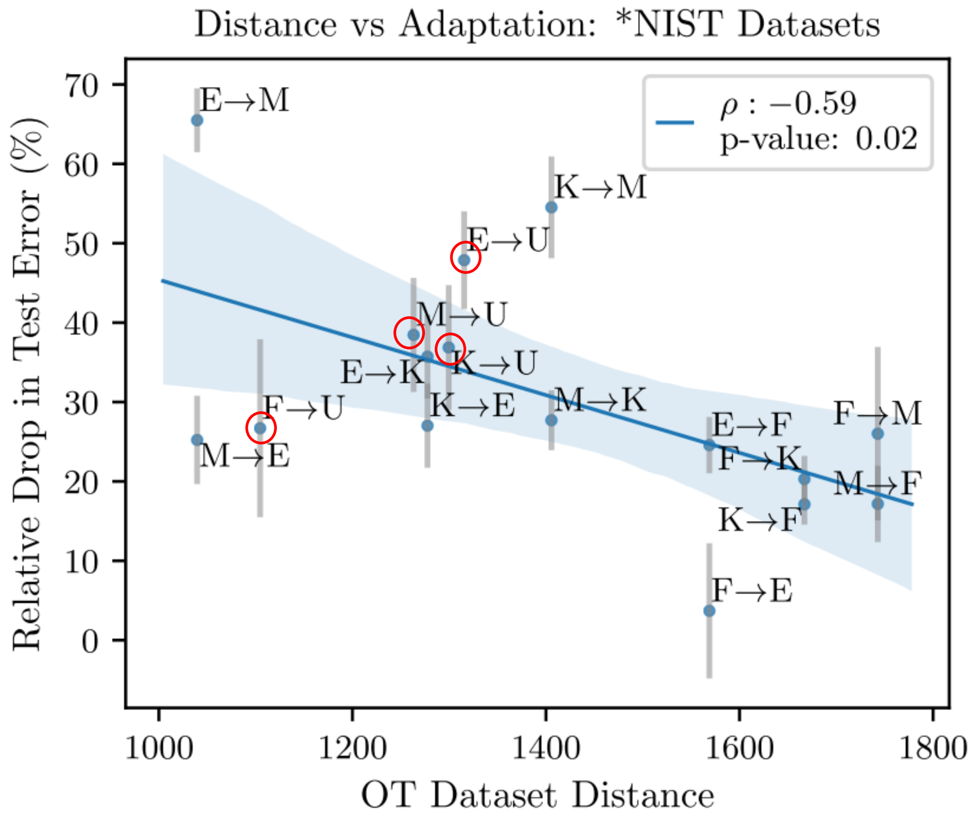


Experiments

- Dataset Selection for Transfer Learning

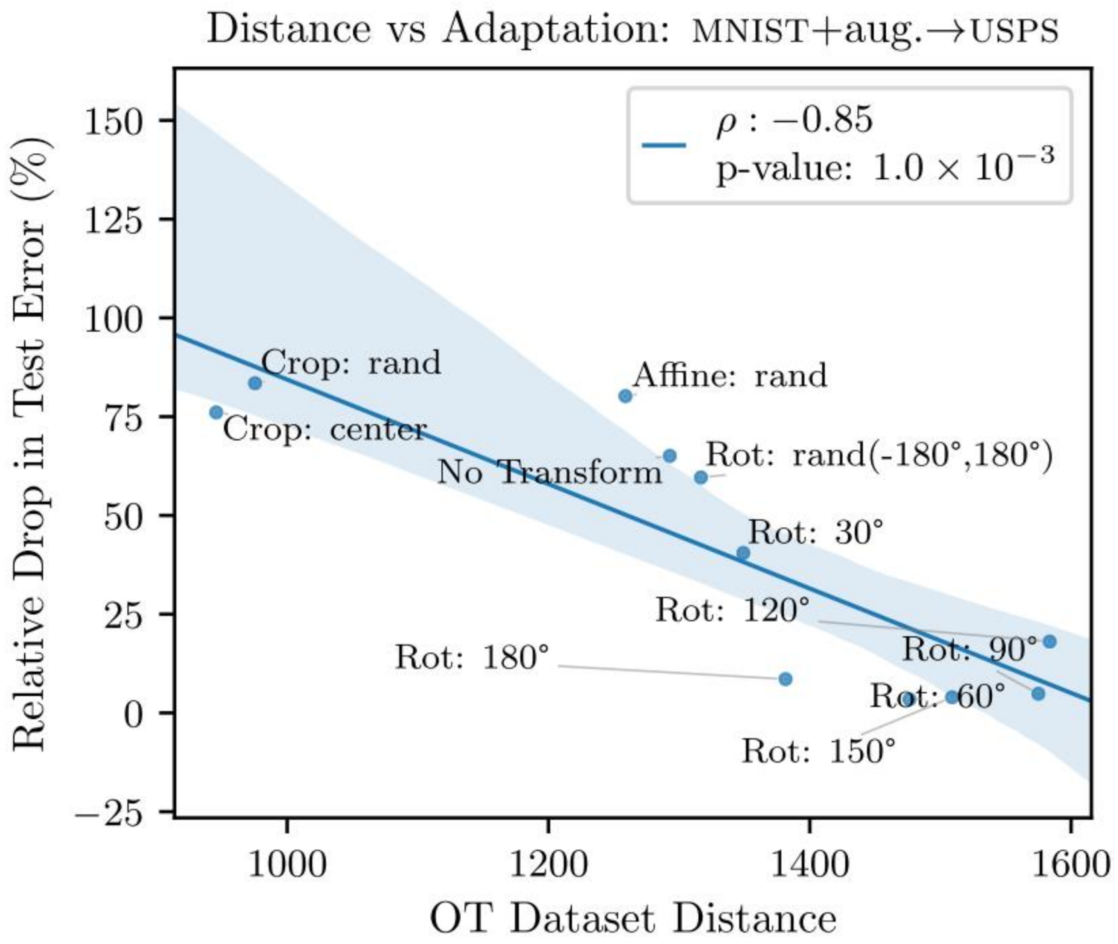
$$\mathcal{T}(\mathcal{D}_S \to \mathcal{D}_T) = 100 \times \frac{\text{error}(\mathcal{D}_S \to \mathcal{D}_T) - \text{error}(\mathcal{D}_T)}{\text{error}(\mathcal{D}_T)}$$

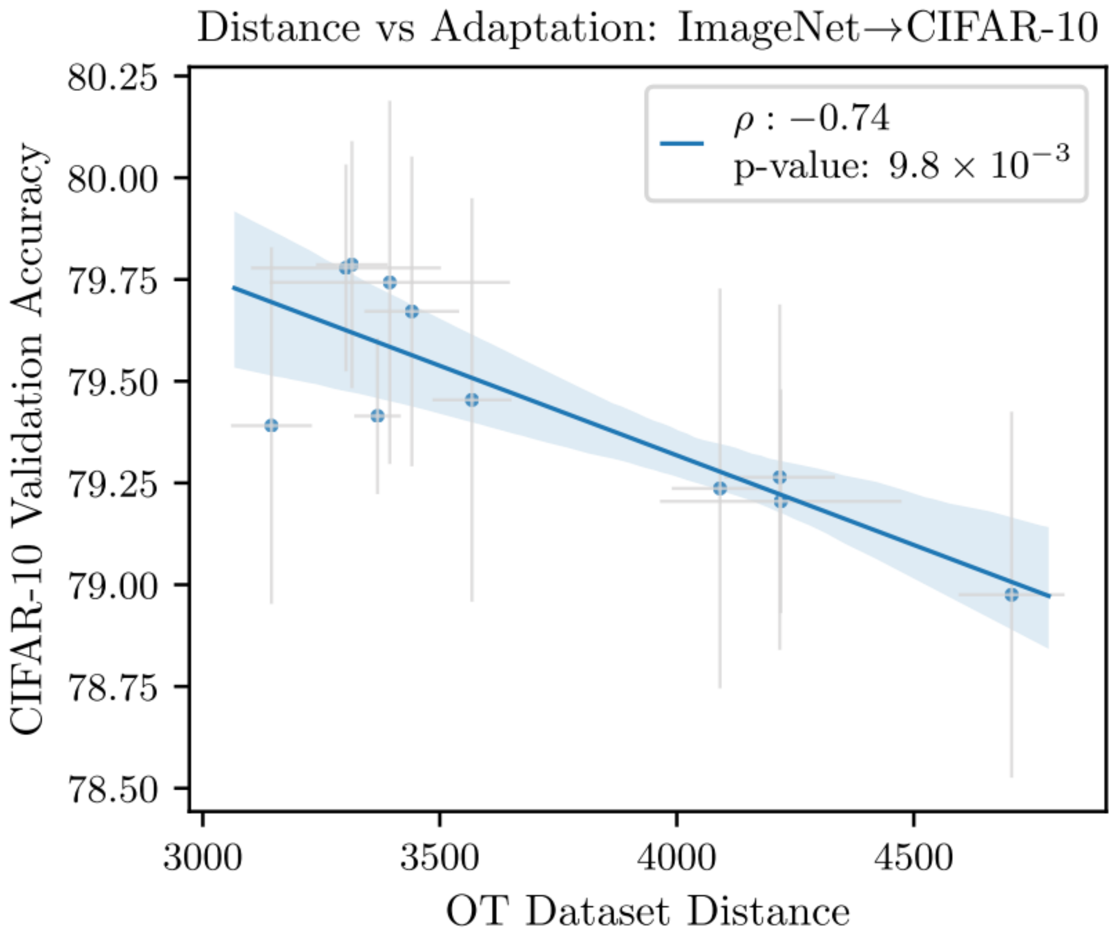




Experiments

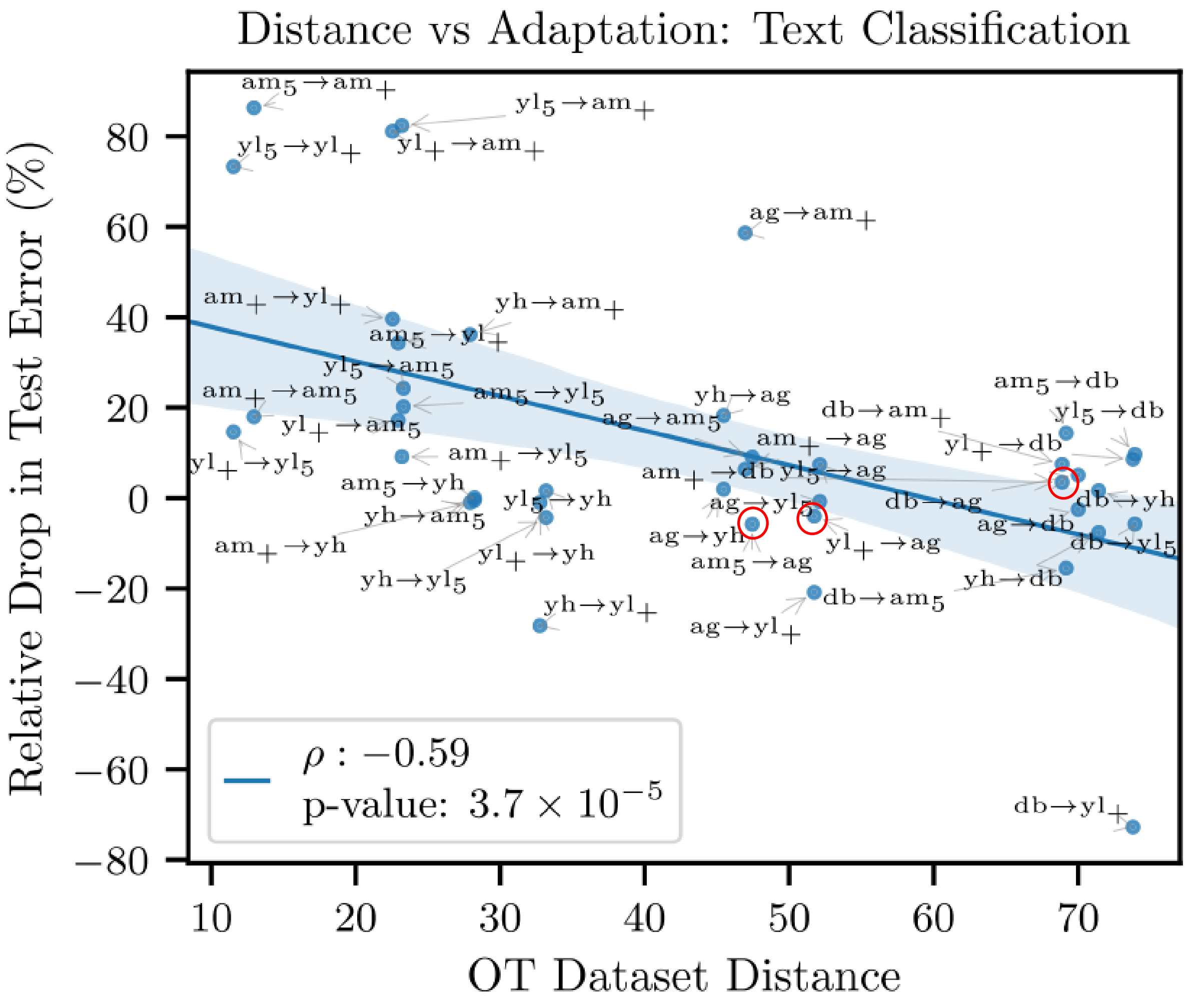
- Distance-Driven Data Augmentation





Experiments

- Transfer Learning for Text Classification



Discussion

- This paper proposes a distance metric based on Optimal Transport to measure the distance between two datasets considering both point-to-point and label-to-label correspondences.
- They did not show the experimental comparisons with other metrics, e.g. KL divergence, and we want to know whether this metric has better consistency to reveal transferability than other methods.