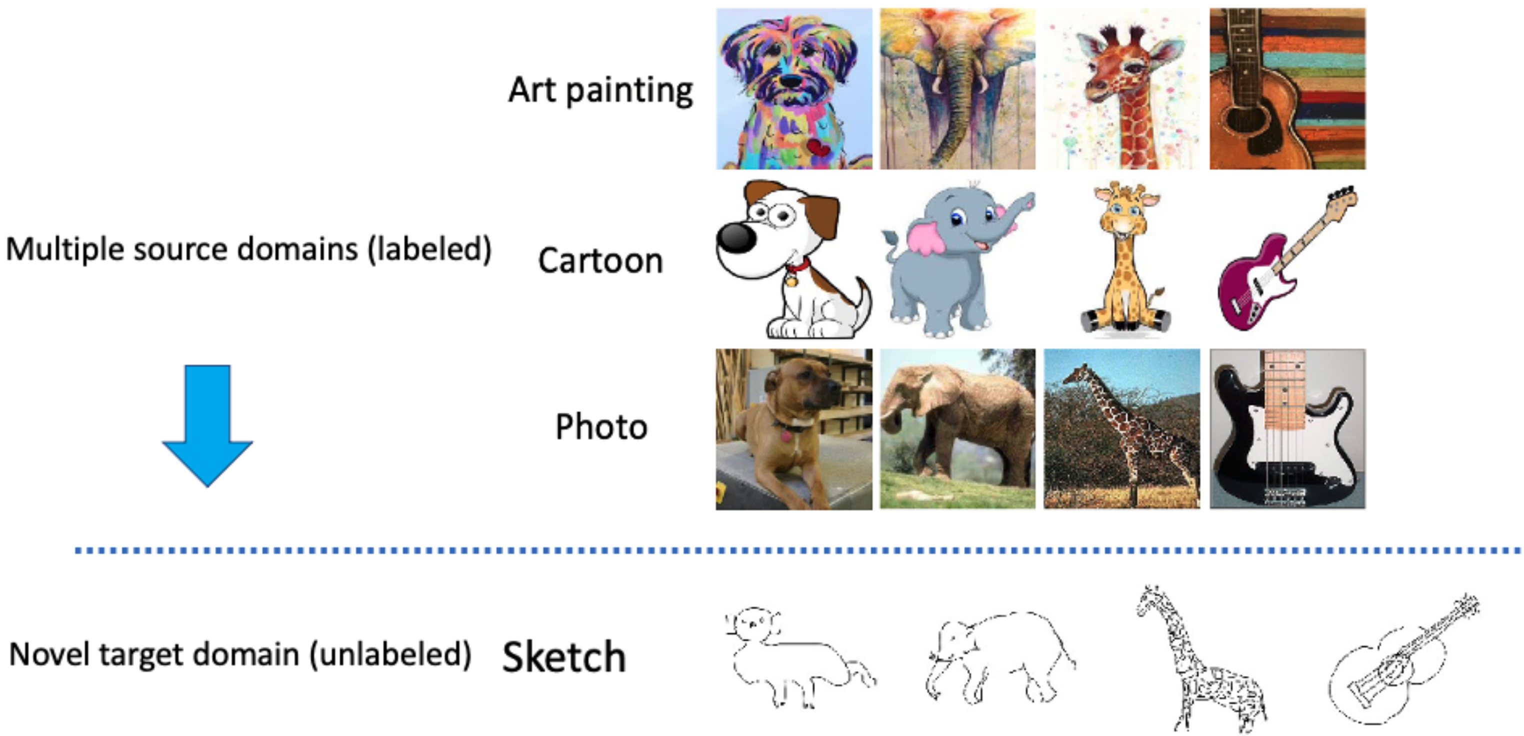
Partial Disentanglement for Domain Adaptation

Lingjing Kong[1], Shaoan Xie[1], Weiran Yao[1], Yujia Zheng[1], Guangyi Chen[2,1], Petar Stojanov[3], Victor Akinwande[1], Kun Zhang[2,1]

[1]Carnegie Mellon University

[2]Mohamed bin Zayed University of Artificial Intelligence

[3]Broad Institute of MIT and Harvard



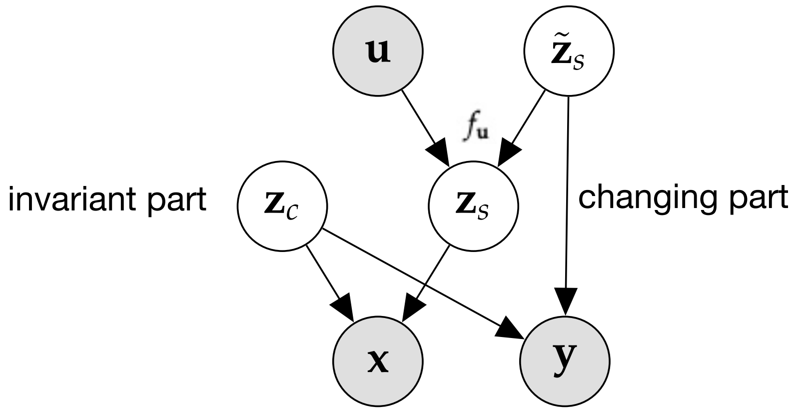
Multi-source Domain Adaptation:

- Resources: labeled data $(\mathbf{x}^{(i)}, y^{(i)})$ for source domains $i = 1, \ldots,$ and unlabeled data $\mathbf{x}^{(\tau)}$ for the target domain τ.
- Goal: learning a strong classifier $p_{y|\mathbf{x},\tau}$ for the target domain τ.

Ill-posedness: $p_{\mathbf{x}|\tau} \overset{????}{\Longrightarrow} p_{\mathbf{x},y|\tau}$

Our contribution:

- We formulate the multi-source domain adaptation problem in the form of a *latent variable model* in light of the *minimal change principle*.
- Under mild assumptions, we show that the latent space is *partial identifiable*.
- Based on the theoretical insight, we propose a practical approach consisting of VAE and flow architectures.



$$\mathbf{z}_c \sim p_{\mathbf{z}_c},\ \tilde{\mathbf{z}}_s \sim p_{\tilde{\mathbf{z}}_s},\ \mathbf{z}_s = f_{\mathbf{u}}(\tilde{\mathbf{z}}_s),\ \mathbf{x} = g(\mathbf{z}_c, \mathbf{z}_s).$$

- Partitioned latent space: the invariant part $\mathbf{z}_c$ and the changing part $\mathbf{z}_s$.
- Minimal change: the domain influence function $f_{\mathbf{u}}$ being component-wise monotonic.

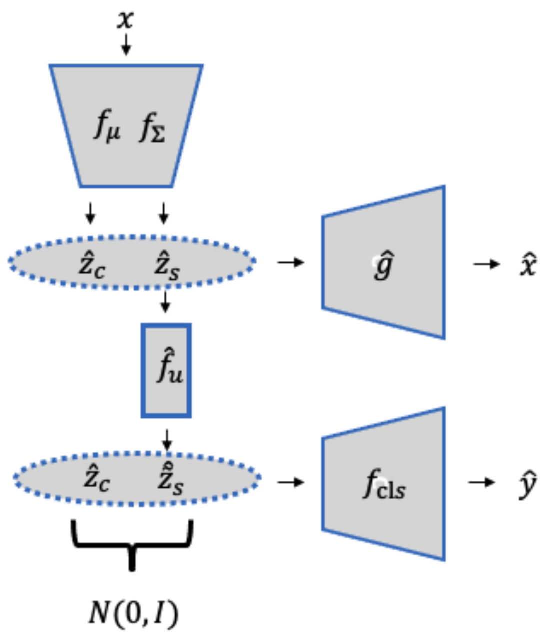
Domain adaptation $\Longrightarrow$ how to identify $(\mathbf{z}_c, \tilde{\mathbf{z}}_s)$ from unlabeled data $(\mathbf{x}, \mathbf{u})$?

Theorem 1

(Informal) Under the assumed data generating process and additional assumptions (e.g. sufficient variability of $p_{\mathbf{z}_s|\mathbf{u}}$ over domains), $\mathbf{z}_s$ and $\mathbf{z}_c$ can be recovered up to component-wise indeterminacy and block-wise indeterminacy respectively.

- Therefore, we can estimate the true latent variables $(\mathbf{z}_c, \mathbf{z}_s)$ from unlabeled data $(\mathbf{x}, \mathbf{u})$.
- Further, we can recover $(\mathbf{z}_c, \tilde{\mathbf{z}}_s)$ and learn a classifier $p_{y|\mathbf{z}_c,\tilde{\mathbf{z}}_s}$ that is applicable to *all domains*.

Proposed Architecture: iMSDA



$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{ent}}.$$

- $\mathcal{L}_{\text{VAE}}$: VAE $(f_\mu, f_\Sigma, \hat{g})$ and flow $(\hat{f}_{\mathbf{u}})$ are trained to estimate the joint distribution $p_{\mathbf{x}, \mathbf{z}_c, \tilde{\mathbf{z}}_s | \mathbf{u}}$.
- $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ent}}$: cross-entropy $\mathcal{L}_{\text{cls}}$ on source domains and conditional entropy $\mathcal{L}_{\text{ent}}$ on the target domain are used to train a classifier (f_{cls}) to estimate $p_{y|\mathbf{z}_c, \tilde{\mathbf{z}}_s}$.

Methods	$\rightarrow$ Art	$\rightarrow$ Cartoon	$\rightarrow$ Photo	$\rightarrow$ Sketch	Avg
Source Only	74.9 $\pm$ 0.88	72.1$\pm$0.75	94.5$\pm$0.58	64.7$\pm$1.53	76.6
DANN	81.9$\pm$1.13	77.5$\pm$1.26	91.8$\pm$1.21	74.6$\pm$1.03	81.5
CMSS	88.6 $\pm$0.36	90.4$\pm$ 0.80	96.9$\pm$0.27	82.0$\pm$0.59	89.5
LtC-MSDA	90.19	90.47	97.23	81.53	89.8
T-SVDNet	90.43	90.61	**98.50**	85.49	91.25
iMSDA (Ours)	$\mathbf{93.75} \pm 0.32$	$\mathbf{92.46} \pm 0.23$	98.48 ± 0.07	$\mathbf{89.22} \pm 0.73$	**93.48**

Table: Classification results on PACS. We employ Resnet-18 as our encoder backbone. We choose $\alpha_1 = 0.1$ and $\alpha_2 = 5e - 5$. The latent space is partitioned with $n_s = 4$ and $n = 64$.

- On multiple benchmark datasets (e.g. PACS), our approach achieves superior perform over all transfer directions.

Thank You!