Deep Generative and Discriminative Domain Adaptation

Han Zhao[†], Junjie Hu[†], Zhenyao Zhu[◇], Adam Coates[♭], & Geoffrey J. Gordon[†]

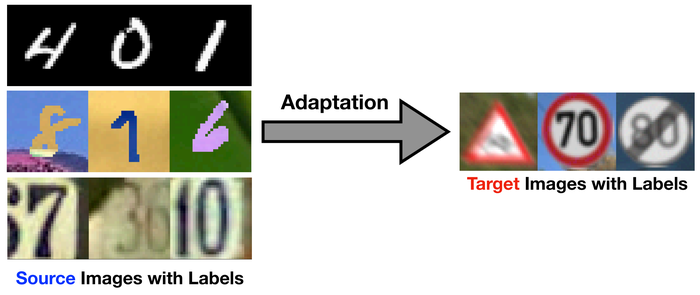
[†]**Carnegie Mellon University,** [◇]**Google,** [♭] **Apple**

{han.zhao,junjieh,ggordon}@cs.cmu.edu, zhuzychn@gmail.com, acoates@cs.stanford.ed

MACHINE LEARNING DEPARTMENT

Summary

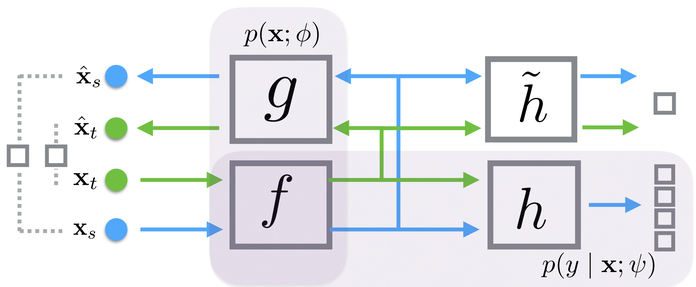
Unsupervised Domain adaptation: Source $\neq$ Target



- **Q:** What if we also have unlabeled data from source domain? Can we utilize them? **A:** By generative modeling.
- **Q:** Under what assumptions can we expect unsupervised domain adaptation succeed? **A:** Matching marginal data distribution.
- **Q:** Is it possible to extend domain adaptation to time-series modeling? **A:** Frame-wise extension.

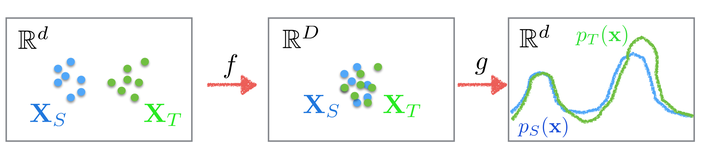
Our Approach

Domain adversarial **Auto**-encoder (DAuto):



Three components:

- Autoencoder: Encoder function f + decoder function g.
- Adversarial discriminator: f + binary classifier $\hat{h}$.
- Target classifier: f + classifier/regressor h.



Carnegie Mellon University

Formulation and Analysis

Joint likelihood maximization:

$$\max_{\psi,\phi} \sum_{i=1}^{m} \log p(y_i \mid \mathbf{x}_i;\psi) + \lambda \sum_{j=1}^{n} \underbrace{-||\mathbf{x}_j - g(f(\mathbf{x}_j;\zeta);\phi\backslash\zeta)||_2^2}_{\text{Lower bound of } \log p(\mathbf{x})}$$

- $p(\mathbf{x})$ given by kernel density estimation

Overall objective function:

$$\min_{W_f,W_g,W_h} \max_{W_d} \quad \sum_{i=1}^{m} \mathcal{L}_y(\mathbf{x}_i, y_i; W_f, W_h) + \lambda \sum_{j=1}^{n} \mathcal{L}_r(\mathbf{x}_j; W_f, W_g) - \mu \sum_{j=1}^{n} \mathcal{L}_d(\mathbf{x}_j; W_f, W_d)$$

- $\mathcal{L}_y$: classification/regression loss
- $\mathcal{L}_r$: reconstruction loss
- $\mathcal{L}_d$: binary classification loss from domain classifier

Analysis: With probability $\geq 1-\delta, \forall h$,

$$\text{err}_T(h) \leq \widehat{\text{err}}_S(h) + \frac{1}{2} d_{\tilde{\mathcal{H}}\Delta\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + \lambda + \frac{10r\Lambda}{c} + \tilde{O}\left(\sqrt{\frac{\log(1/\delta)}{m}}\right)$$

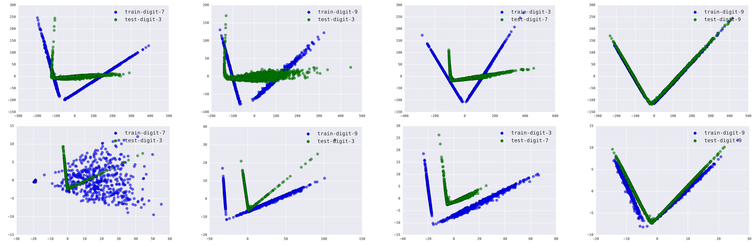
- $d_{\mathcal{H}}(\cdot,\cdot)$ measures distance between two distributions
- λ = the optimal classification error achievable in both domains
- $r^2 := \Sigma_{i=1}^{m} ||\mathbf{x}_i - g(f(\mathbf{x}_i))||_2^2/m$ the average reconstruction error

Experiments

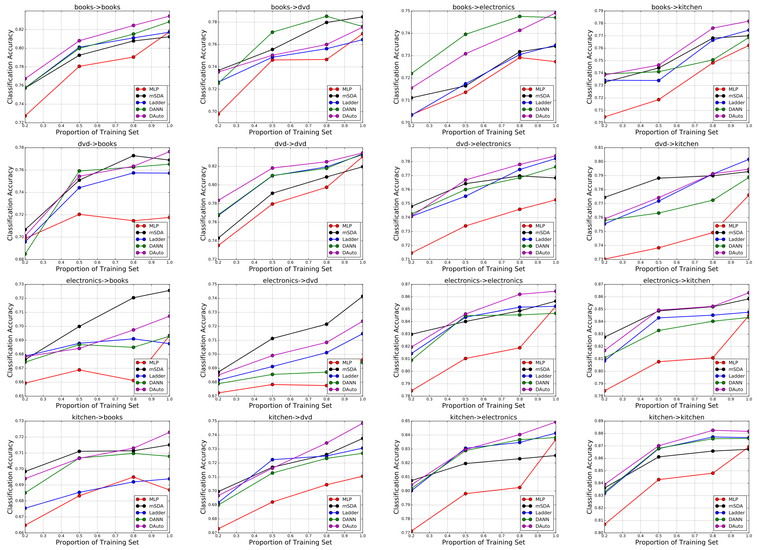
Datasets (Train/Test):

- Image: 10 digit classification
  - MNIST: 60,000/10,000
  - SVHN: 73,257/26,032
  - USPS: 7,291/2,007
- Text: sentiment analysis
  - Books (B): 2,000/4,465
  - DVDs (D): 2,000/3,586
  - Electronics (E): 2,000/5,681
  - Kitchen appliances (K): 2,000/5,945
- Speech: speech recognition
  - Native: $\sim$ 25/7 hours
  - Chinese accent: $\sim$ 25/7 hours
  - Indian accent: $\sim$ 25/7 hours

Visualization on Synthetic Experiments: DAuto aligns features from both source and target domains.



Sentiment Analysis: Unlabeled data/Semi-supervised Learning Helps.



Digit Classification: Works in multi-class setting.

	No Adapt			DANN		
	SVHN	**MNIST**	**USPS**	**SVHN**	**MNIST**	**USPS**
SVHN	0.8553	0.5459	0.5277	0.8596	0.5690	0.5426
MNIST	0.2054	0.9883	0.6442	**0.2241**	0.9880	0.6500
USPS	0.1628	0.3396	0.9507	0.1585	0.3562	0.9517
	ADDA			**DAuto**		
	SVHN	**MNIST**	**USPS**	**SVHN**	**MNIST**	**USPS**
SVHN	**0.8707**	0.5542	0.5561	0.8626	**0.5864**	**0.5655**
MNIST	0.2091	**0.9894**	**0.6856**	0.2086	0.9869	0.6428
USPS	0.1602	0.3570	0.9512	**0.1717**	**0.3762**	**0.9537**

Speech Recognition: Improved results between Native and Chinese.