



Department of Biostatistics & Bioinformatics

Duke University School of Medicine

Adversarial Discriminative Domain Adaptation

Machine Learning in Practice Reading Group

Duke B&B

February 20, 2023

Presented by Yuqi (Lucilla) Li

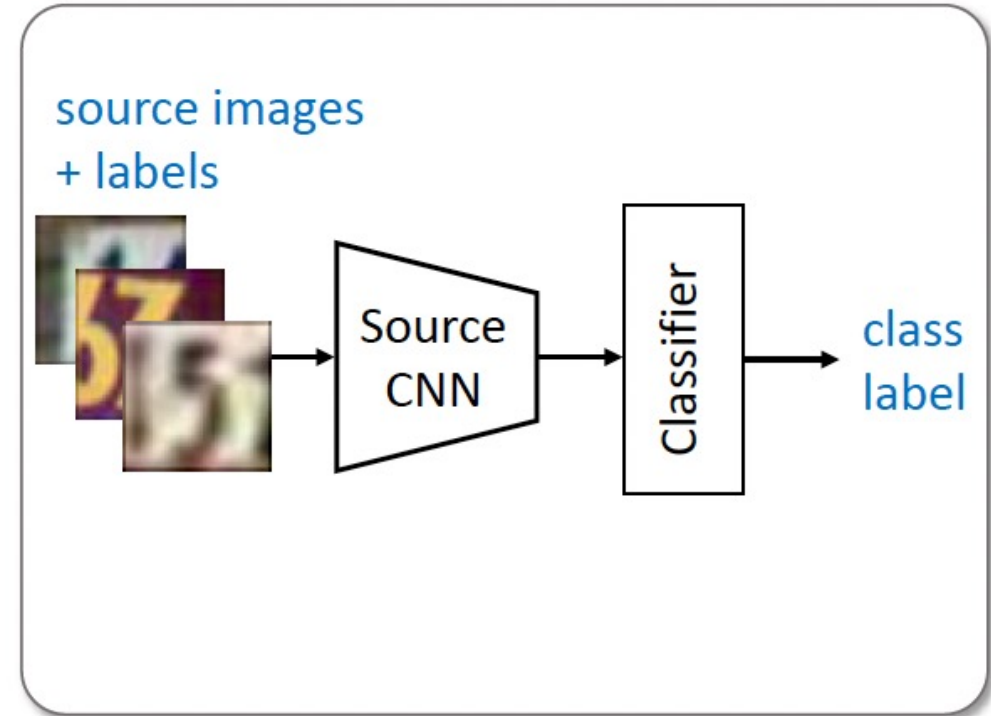
1. Introduction and Background

Example: what if domain shift?

- Fine-tune the large number of parameters?
- Enough labeled data?

Domain Adaptation Methods

- Map both domains into a common feature space
- Reconstruct the target domain from the source representation



Adversarial Adaptation:

- minimize an approximate domain discrepancy distance through an adversarial objective with respect to a domain discriminator

Achievement of the paper:

1. Generalized Framework for Adversarial Adaptation

- Examine the different factors of variation between existing approaches
- Unifies design choices: weight-sharing, base models, adversarial loss

2. Adversarial Discriminative Domain Adaptation

- unexplored unsupervised adversarial adaptation method (new combination)

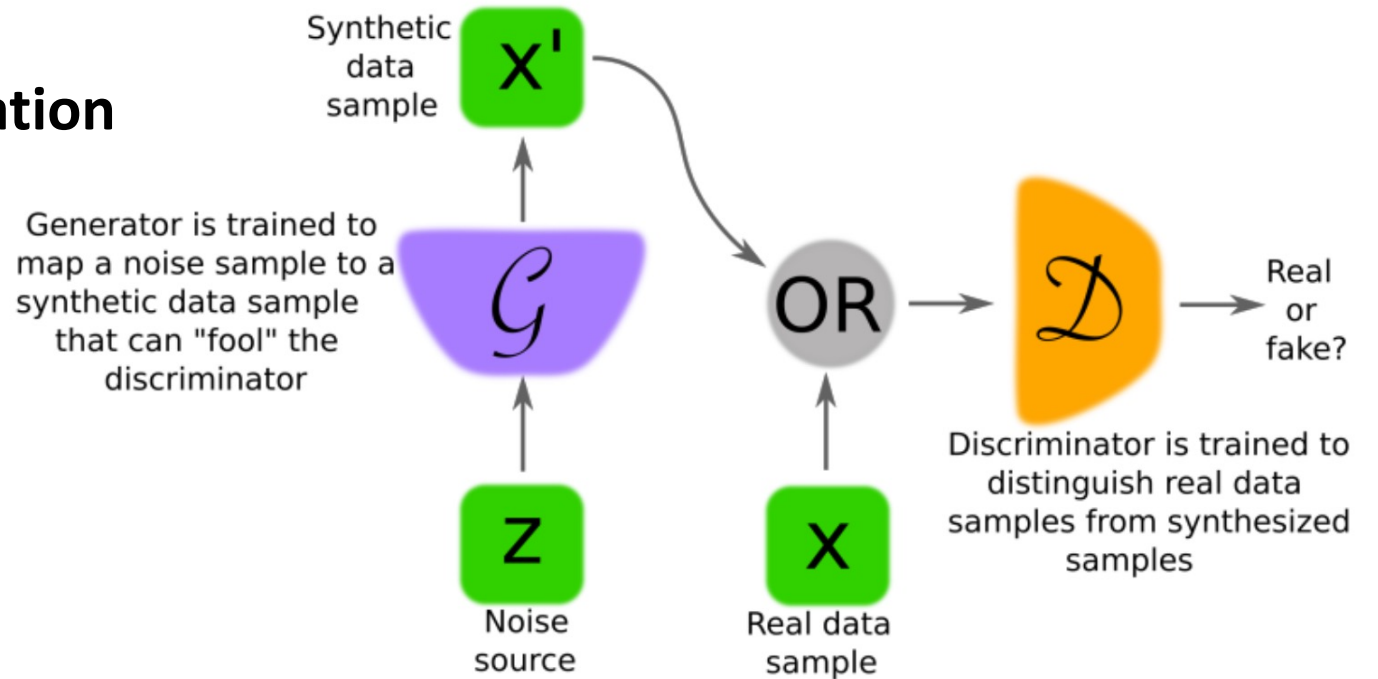
Generative Adversarial Learning:

- Generative Adversarial Network (GAN)
- Generator and Discriminator

Different choices in Domain Adaptation

Algorithms:

- use a generator or not
- which loss function
- share weights or not



2. Related Work

- Transferring deep neural network representations from a labeled source datasets to a target domain where labeled data is sparse or non-existent .
- **Domain Confusion**: adding a domain classifier, designed a domain confusion loss to encourage its prediction to be close to a uniform distribution over binary labels
- **Gradient Reversal**: also treats domain invariance as a binary classification problem, directly maximize the loss of the domain classifier by reversing its gradients
- **CoGAN**: training two GANs to generate the source and target images respectively, tying the high-level layer parameters of the two GANs, same noise input can generate a corresponding pair of images from the two distributions

3. Generalized Framework for Adversarial Adaptation

- In unsupervised adaptation, we assume access to source \mathbf{X}_s and labels Y_s drawn from a source domain distribution $p_s(x, y)$, as well as target \mathbf{X}_t drawn from a target distribution $p_t(x, y)$
- **Goal: learn a target representation M_t and classifier C_t that can classify target**
- Domain adaptation learns a source representation mapping M_s and classifier C_s , then learns to adapt that model for use in the target domain
- Regularize the learning of M_s and M_t , to minimize the distance between the empirical source and target mapping distributions: $M_s(\mathbf{X}_s)$ and $M_t(\mathbf{X}_t)$
- If it is the case, the source classification model C_s can be directly applied to the target representations, setting $C = C_s = C_t$, using standard supervised loss.

How to minimize the source and target representation distances?

- Alternating minimization between two functions: constraints, adversarial loss.
- Domain discriminator D , which classifies whether a data point is drawn from the source or the target domain. D is optimized according to a standard supervised loss, where the labels indicate the origin domain:

$$\mathcal{L}_{\text{adv}_D}(\mathbf{X}_s, \mathbf{X}_t, M_s, M_t) = -\mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s} [\log D(M_s(\mathbf{x}_s))] - \mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t} [\log(1 - D(M_t(\mathbf{x}_t)))]$$

- **The source and target mappings** are optimized according to a **constrained** adversarial objective (vary across methods).
- Generic formulation:
$$\begin{aligned} \min_D \quad & \mathcal{L}_{\text{adv}_D}(\mathbf{X}_s, \mathbf{X}_t, M_s, M_t) \\ \min_{M_s, M_t} \quad & \mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t, D) \\ \text{s.t.} \quad & \psi(M_s, M_t) \end{aligned}$$

1. Parameterization of Source and Target Mappings

- Determine the mapping parameterization for source. Initialize the target mapping parameters with the source, but different constraints $\psi(M_s, M_t)$
- Make sure the target mapping is set to minimize the distance between the S and T domains, while maintaining a target mapping that is category discriminative

1. A common form of constraint is source and target layerwise equality:

$$\psi_{\ell_i}(M_s^{\ell_i}, M_t^{\ell_i}) = (M_s^{\ell_i} = M_t^{\ell_i})$$

2. All layers are constrained, thus enforcing exact source and target mapping consistency. Symmetric transformation: may make the optimization poorly conditioned.
3. Learn an asymmetric transformation with only a subset of the layers constrained. Partially shared weights can lead to effective adaptation.

2. Adversarial Losses

- The **gradient reversal** layer optimizes the mapping to maximize the discriminator loss directly (corresponds to the **true minimax** objective for GAN):

$$\mathcal{L}_{\text{adv}_M} = -\mathcal{L}_{\text{adv}_D}$$

which can be problematic, since early on during training the discriminator converges quickly, causing the gradient to vanish.

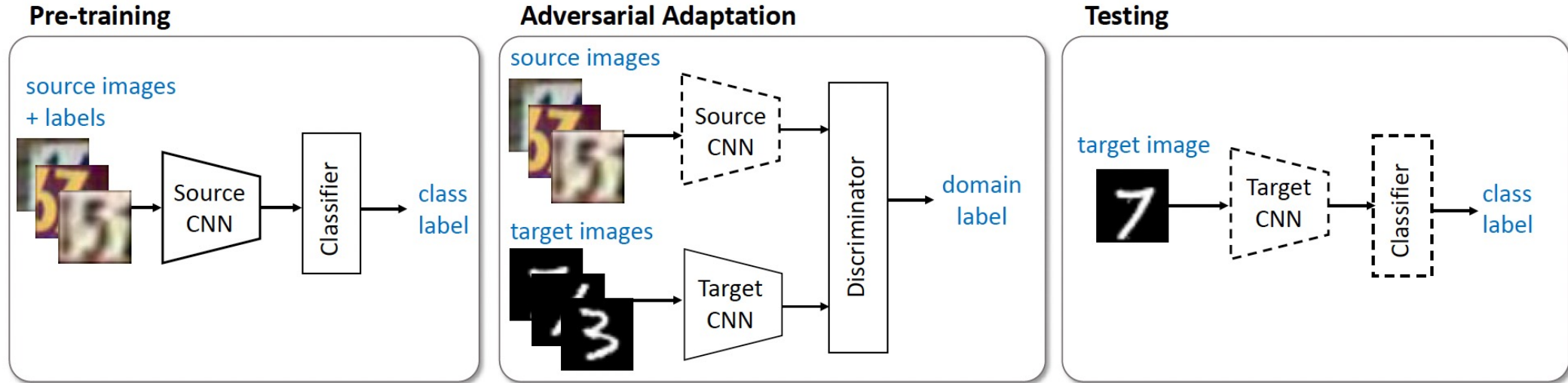
- **GAN-based loss function:**

Train the generator with the standard loss with inverted labels. \mathcal{L}_{adv_D} remains unchanged, but \mathcal{L}_{adv_M} becomes:

$$\mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t, D) = -\mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t} [\log D(M_t(\mathbf{x}_t))].$$

It has the same properties as the minimax loss but provides stronger gradients to the target mapping.

4. Adversarial Discriminative Domain Adaptation



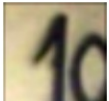


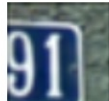
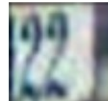
- Discriminative base model: adversarial adaptive methods optimize directly in a discriminative space
- Unshared weights: allow independent source and target mappings, more domain specific feature extraction to be learned
- GAN loss: learning an asymmetric mapping: modify the target model to match the source distribution, until it is indistinguishable, similar to GAN




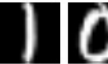














5. Results of Experiments

Method	Base model	Weight sharing	Adversarial loss
Gradient reversal [16]	discriminative	shared	minimax
Domain confusion [12]	discriminative	shared	confusion
CoGAN [13]	generative	unshared	GAN
ADDA (Ours)	discriminative	unshared	GAN

MNIST     

USPS     

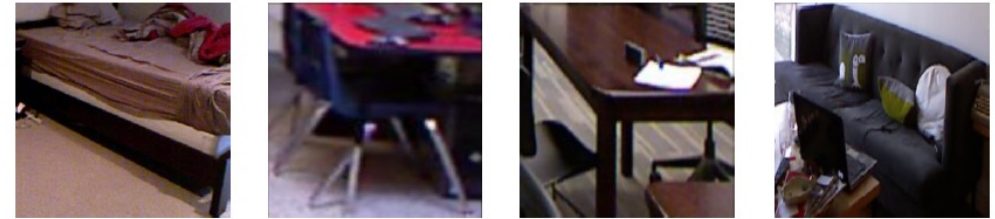
SVHN     

Method	MNIST → USPS    →   	USPS → MNIST    →   	SVHN → MNIST    →   
Source only	0.752 ± 0.016	0.571 ± 0.017	0.601 ± 0.011
Gradient reversal	0.771 ± 0.018	0.730 ± 0.020	0.739 [16]
Domain confusion	0.791 ± 0.005	0.665 ± 0.033	0.681 ± 0.003
CoGAN	0.912 ± 0.008	0.891 ± 0.008	did not converge
ADDA (Ours)	0.894 ± 0.002	0.901 ± 0.008	0.760 ± 0.018

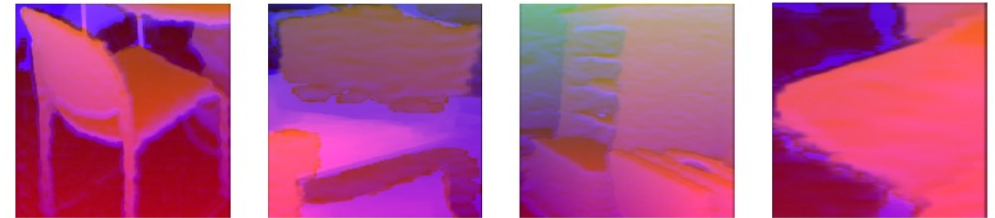
5. Results of Experiments

- Cross-modality adaptation
- Improve the average accuracy from 13.9% to 21.1%

RGB



HHA



	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	television	toilet	overall
# of instances	19	96	87	210	611	103	122	129	25	55	144	37	51	276	47	129	210	33	17	2401
Source only	0.000	0.010	0.011	0.124	0.188	0.029	0.041	0.047	0.000	0.000	0.069	0.000	0.039	0.587	0.000	0.008	0.010	0.000	0.000	0.139
ADDA (Ours)	0.000	0.146	0.046	0.229	0.344	0.447	0.025	0.023	0.000	0.018	0.292	0.081	0.020	0.297	0.021	0.116	0.143	0.091	0.000	0.211
Train on target	0.105	0.531	0.494	0.295	0.619	0.573	0.057	0.636	0.120	0.291	0.576	0.189	0.235	0.630	0.362	0.248	0.357	0.303	0.647	0.468

6. Conclusion Remarks

- The unified framework for unsupervised domain adaptation **provides a cohesive simplified view**, illustrating the similarities and differences between methods, people can benefit from the key ideas and **combine the strategies into new method**
- **ADDA generalized well across various tasks**, additional analysis indicates that the representations learned via ADDA resemble features learned with supervisory data in the target domain much more closely than unadapted features
- ADDA was proposed 5 years ago, there is quite state-of-the-art progress for this question, ADDA paved the path towards new discoveries

Reference

1. Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. <https://doi.org/10.48550/arxiv.1702.05464>
2. Vlachostergiou A, Caridakis G, Mylonas P, Stafylopatis A. Learning Representations of Natural Language Texts with Generative Adversarial Networks at Document, Sentence, and Aspect Level. *Algorithms*. 2018; 11(10):164. <https://doi.org/10.3390/a11100164>

Thanks for listening!