On Learning Invariant Representations for Domain Adaptation

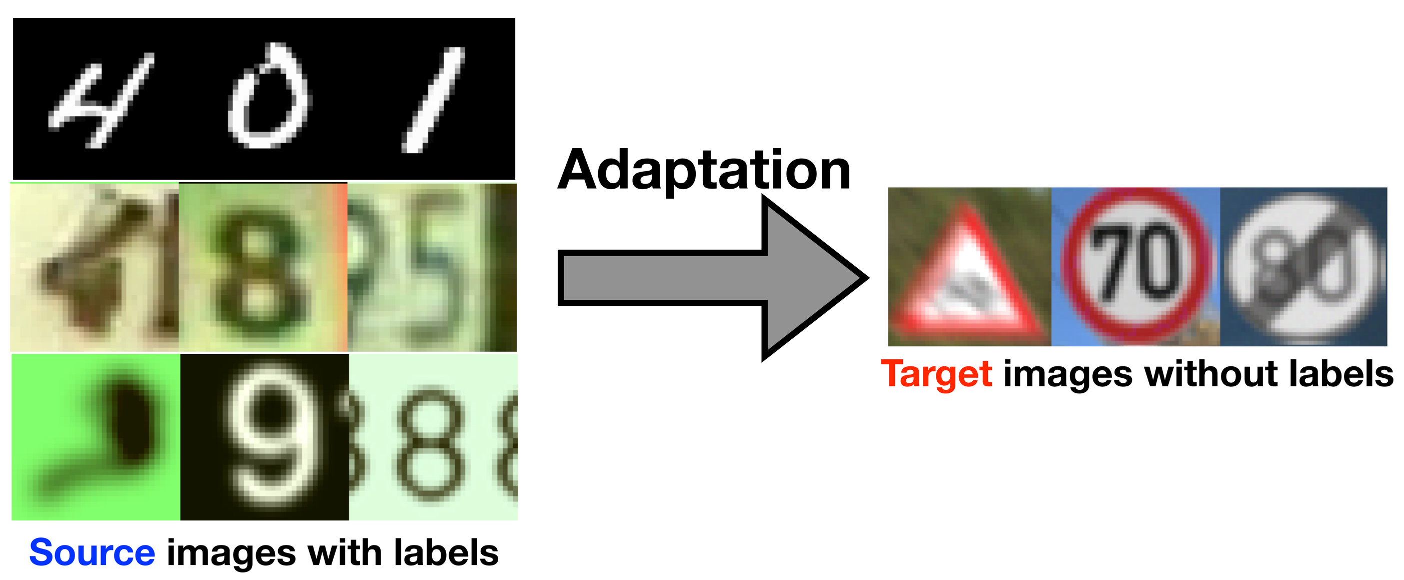
Han Zhao[†], Remi Tachet des Combes[‡], Kun Zhang[†] & Geoffrey J. Gordon[†,‡]

[†]Carnegie Mellon University, [‡]Microsoft Research Montreal

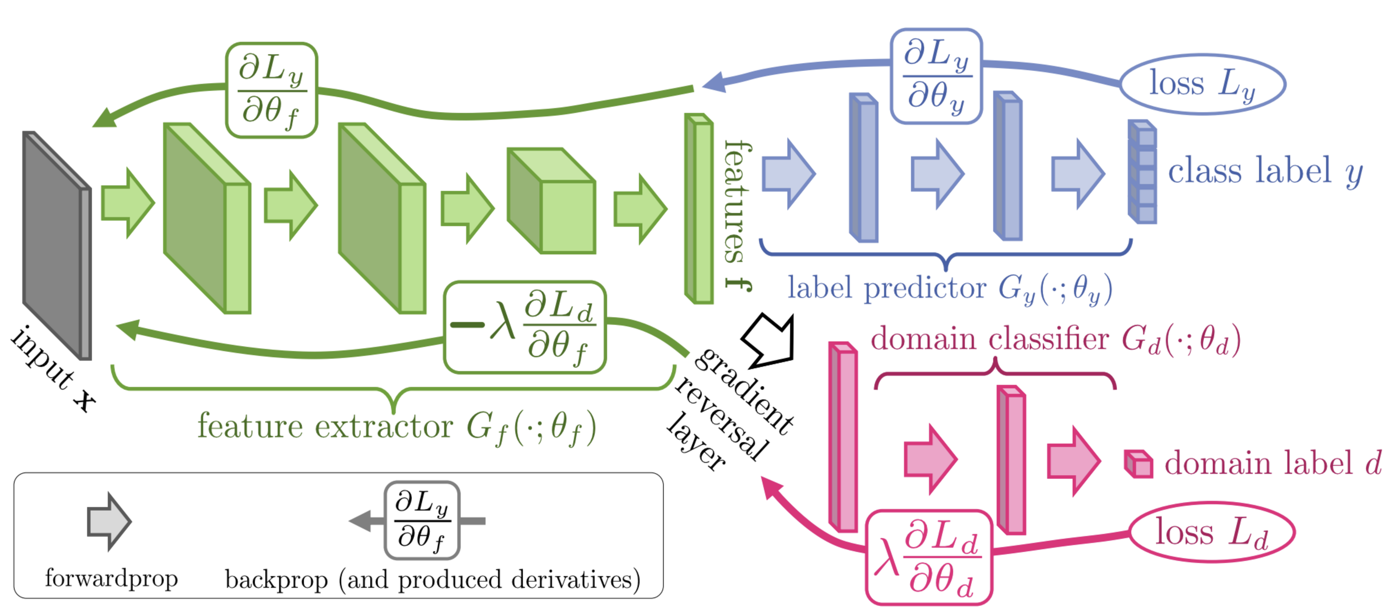
han.zhao@cs.cmu.edu, kunz1@andrew.cmu.edu, {remi.tachet, geoff.gordon}@microsoft.com

Overview

Unsupervised Domain Adaptation: Source $\neq$ Target



Domain Adversarial Neural Network (DANN):



Question:

Is finding invariant representations while at the same time achieving a small source error sufficient to guarantee a small target error? If not, under what conditions is it?

Our Answer: No, and it provably hurts if the label distributions are different! Only sufficient when conditional distributions are close.

Our Contributions:

- A simple counter-example that invalidates domain adaptation algorithms based on matching marginal distributions.
- Sufficient condition: a generalization upper bound that suggests matching conditional distributions.
- Necessary condition: an information-theoretic lower bound that suggests matching marginal label distributions.
- Empirical results that corroborate our theoretical analysis.

Carnegie Mellon University

Microsoft Research

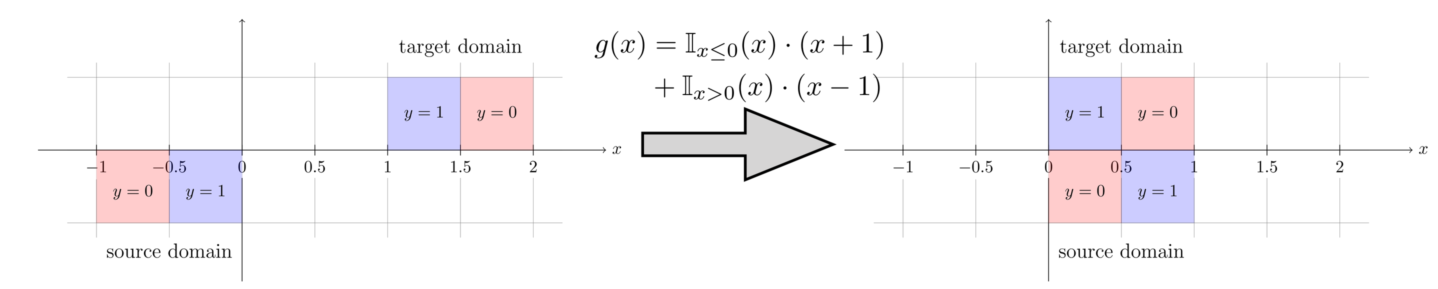
A Simple Example

Previous Theoretical Result: Given hypothesis class $\mathcal{H}$ and $\mathcal{A}_{\mathcal{H}} := \{h^{-1}(1) \mid h \in \mathcal{H}\}$, the **$\mathcal{H}$-divergence** is: $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') := \sup_{A\in\mathcal{A}_{\mathcal{H}}} |\Pr_{\mathcal{D}}(A) - \Pr_{\mathcal{D}'}(A)|$. Generalization bound for binary classification (Blitzer et al. NeurIPS' 08), $\forall h \in \mathcal{H}$:

$$\varepsilon_T(h) \leq \underbrace{\varepsilon_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)}_{\text{To be minimized by learning invariant representations}} + \lambda^*$$

- $\varepsilon_S(h)/\varepsilon_T(h)$: population source/target binary classification error.
- $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$: divergence between source and target domains.
- $\lambda^* := \min_{h'\in\mathcal{H}} \varepsilon_S(h') + \varepsilon_T(h')$, the optimal joint error.

A counter-example:



Before adaptation:

$$\mathcal{D}_S = U(-1, 0), \qquad f_S(x) = \begin{cases} 0, & x \leq -1/2 \\ 1, & x > -1/2 \end{cases}$$

$$\mathcal{D}_T = U(1, 2), \qquad f_T(x) = \begin{cases} 0, & x \geq 3/2 \\ 1, & x < 3/2 \end{cases}$$

Optimal hypothesis: $h^*(x) = 1$ iff $x \in (-1/2, 3/2)$ with $\lambda^* = 0$.

Feature transformation:

$$g(x) = \mathbb{I}_{x\leq 0}(x)\cdot(x+1) + \mathbb{I}_{x>0}(x)\cdot(x-1)$$

After adaptation: Perfect domain alignment: $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_S, \mathcal{D}'_T) = 0$. But,

$$\mathcal{D}'_S = U(0, 1), \qquad f'_S(x) = \begin{cases} 0, & x \leq 1/2 \\ 1, & x > 1/2 \end{cases}$$

$$\mathcal{D}'_T = U(0, 1), \qquad f'_T(x) = \begin{cases} 0, & x \geq 1/2 \\ 1, & x < 1/2 \end{cases}$$

Now $\forall h' : \mathbb{R} \mapsto \{0, 1\}$, $\quad \varepsilon_S(h' \circ g) + \varepsilon_T(h' \circ g) = 1$, hence $\lambda'^* = 1$.

Implication: In this example, minimizing the source error while aligning domains will only increase the target error.

Theoretical Analysis

Our generalization upper bound on the target error: Let $\langle\mathcal{D}_S, f_S\rangle$ and $\langle\mathcal{D}_T, f_T\rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\},$$

where $\tilde{\mathcal{H}} := \{\mathrm{sgn}(|h(\mathbf{x}) - h'(\mathbf{x})| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$.

- Free of the unavailable term $\lambda^* := \min_{h'\in\mathcal{H}} \varepsilon_S(h') + \varepsilon_T(h')$.
- Incorporate the conditional shift $\min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$ into the analysis.
- Also holds for (bounded) regression problems.

Our information-theoretic lower bound on the joint error: Suppose $X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$ holds, and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then:

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2.$$

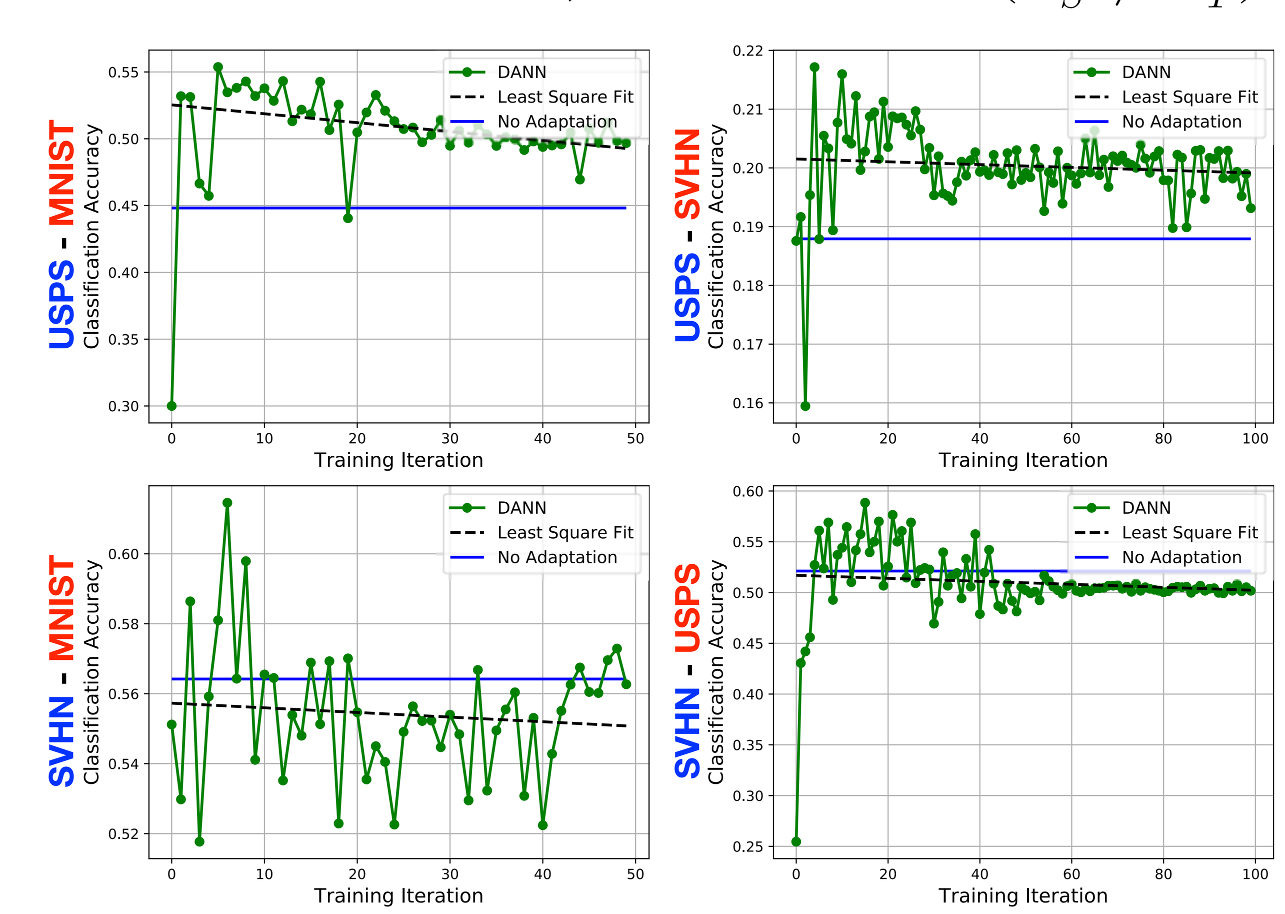
- $d_{\mathrm{JS}}(\cdot,\cdot)$: Jensen-Shannon distance between distributions.
- $\mathcal{D}_S^Y/\mathcal{D}_T^Y$: source/target marginal label distributions.

Extension: Different transformations for the S/T domains don't help: Let g_S, g_T be the source and target transformation functions from $\mathcal{X}$ to $\mathcal{Z}$. Suppose $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then:

$$\varepsilon_S(h \circ g_S) + \varepsilon_T(h \circ g_T) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2.$$

Experiments

Digit classification on MNIST, USPS and SVHN ($\mathcal{D}_S^Y \neq \mathcal{D}_T^Y$):



Conclusion: When the label distributions are different, aligning both domains while minimizing the source error leads to an increasing target error during training.