On Learning Invariant Representations for Domain Adaptation

(2019)

Han Zhao, Remi Tachet des Combes, Kun Zhang, Geoffrey J. Gordon

Summary

Contributions

Due to the ability of deep neural nets to learn rich feature representations, recent advances in domain adaptation have focused on using these networks to learn invariant representations, i.e., intermediate features whose distribution is the same in source and target domains, while at the same time achieving small error on the source domain. The hope is that the learned intermediate representation, together with the hypothesis learned using labeled data from the source domain, can generalize to the target domain. But is this sufficient? is finding invariant representations with a small error on source is enough to get a small target error? The authors show that having invariant representations may lead to high joint error, since learning invariant representation does suppress any domain information that might be helpful when doing classification (in case of conditional shift). To understand this, the authors propose a generalization upper bound as a sufficient condition that explicitly takes into account the conditional shift between source and target domains.

Notations & Preliminaries

- $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ denote the input, output and representation spaces.
- A feature extractor $g : \mathcal{X} \mapsto \mathcal{Z}$.
- A hypothesis $h : \mathcal{X} \to \{0, 1\}$.
- A domain corresponds to a distribution $\mathcal{D}$ and a labeling function $f : \mathcal{X} \to [0, 1]$.
- The ground truths are given by a deterministic labeling function f: $Y = f(X)$.
- The error on the source corresponds to the disagreement between the labeling function and the hypothesis h: $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_S}[|h(\mathbf{x}) - f(\mathbf{x})\|$. In case of binary classification: $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_S}[\mathbb{I}(f(\mathbf{x}) \neq h(\mathbf{x}))]$.
- The risk of hypothesis h is the error of h w.r.t. the true labeling function under the source domain domain $\varepsilon_S(h) := \varepsilon_S(h, f_S)$.
- $\widehat{\varepsilon}_S(h)$ and $\widehat{\varepsilon}_T(h)$ denote the empirical risks on source and target domains.
- The problem of domain adaptation considered in this work can be stated as: under what conditions and by what algorithms can we guarantee that a small training error $\widehat{\varepsilon}_S(h)$ implies a small test error $\widehat{\varepsilon}_T(h)$? Clearly, this goal is not always possible if the source and target domains are far away from each other.
- $\mathcal{H}$-divergence: For hypothesis $\mathcal{H}$ on $\mathcal{X}$ and a collection of subsets $\mathcal{A}_\mathcal{H}$ of $\mathcal{X}$ that are the support of some hypothesis h in $\mathcal{H}$. The distance between the two distributions $\mathcal{D}$ and $\mathcal{D}'$ is $d_\mathcal{H}(\mathcal{D}, \mathcal{D}') := \sup_{A\in\mathcal{A}_\mathcal{H}} |\Pr_\mathcal{D}(A) - \Pr_{\mathcal{D}'}(A)|$
- The joint risk, let a hypothesis $h^* := \arg\min_{h\in\mathcal{H}} \varepsilon_S(h) + \varepsilon_T(h)$ that gives us the minimum joint risk on both the source and target domains, the joint risk of h^* is denotes as $\lambda^* := \varepsilon_S(h^*) + \varepsilon_T(h^*)$

- Bound on the target risk: $\varepsilon_T(h) \leq \widehat{\varepsilon}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T\right) + \lambda^*$. To minimize the upper bound, we need to learn a rich feature extractor g such that the induced target and source distributions on $\mathcal{Z}$ are close: $d_{\mathcal{H}}\left(\mathcal{D}_S^g, \mathcal{D}_T^g\right) = 0$, while still being able to achieve a small empirical error on the source domain with a classifier trained on $\mathcal{Z}$

Method

Is finding invariant representations alone a sufficient condition for the success of domain adaptation?

Counter-example

Let $\mathcal{X} = \mathcal{Z} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$, and let the source and targe distribution be as follows:

$$\mathcal{D}_S = U(-1, 0), \quad f_S(x) = \begin{cases} 0, & x \leq -1/2 \\ 1, & x > -1/2 \end{cases}$$
$$\mathcal{D}_T = U(1, 2), \quad f_T(x) = \begin{cases} 0, & x \geq 3/2 \\ 1, & x < 3/2 \end{cases}$$

In the example above, a hypothesis $h^*(x) = 1$ iff $x \in (-1/2, 3/2)$ gives the perfect optimal performances on both domain. But if we apply a transformation g to align $\mathcal{D}_S$ and $\mathcal{D}_T$, such as:

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

A new hypothesis trained on $g(X)$ will make an error in one of the two domain, as illustrated in the figure bellow.

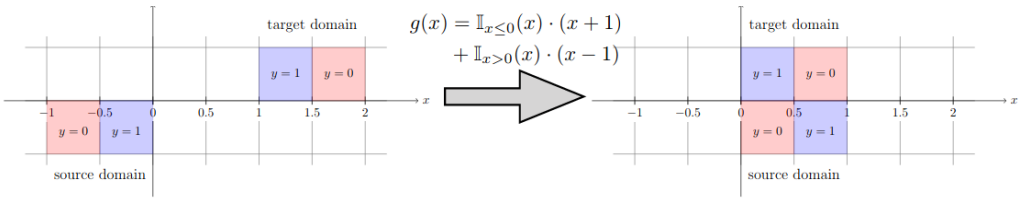

Figure 1. A counterexample where invariant representations lead to large joint error on source and target domains. Before transformation of $g(\cdot)$, $h^*(x) = 1$ iff $x \in (-1/2, 3/2)$ achieves perfect classification on both domains. After transformation, source and target distributions are perfectly aligned, but no hypothesis can achieve a small joint error.

This contradiction comes from the large joint risk λ^*, even if we have a good classifier on source and a small discrepancy. We can also explain this contradiction due to the different labeling functions on the target and the source domains in the induced representations:

$$f'_S(x) = \begin{cases} 0, & x \leq 1/2 \\ 1, & x > 1/2 \end{cases}, \quad f'_T(x) = \begin{cases} 0, & x > 1/2 \\ 1, & x \leq 1/2 \end{cases}$$

Having two far away labeling functions does give us a high joint risk.

A new-bound

The problems with the traditional bound, is that it is intractable to compute the joint risk λ^*, and given that λ^* depend on both risks, on the target and source, the bound in very conservative and loose in many case.

Let and $f_S : \mathcal{X} \to [0, 1]$ be the optimal labeling functions on the source and target domains, respectively. The authors propose the following bound

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

The three terms in the upper bound have natural interpretations: the first term is the source error, the second one corresponds to the discrepancy between the marginal distributions, and the third

measures the distance between the labeling functions from the source and target domains.Altogether, they form a sufficient condition for the success of domain adaptation: besides a small source error, not only do the marginal distributions need to be close, but so do the labeling functions. In this bound, the last term does not depend on the choice of the hypothesis, but the joint risk does, making it more tight.

In the covariate shift setting, where we assume the conditional distributions of $Y|X$ between the source and target domains are the same, the third term in the upper bound vanishes. In that case the above bound guarantees successful domain adaptation, it suffices to match the marginal distributions while achieving small error on the source domain. In general settings where the optimal labeling functions of the source and target domains differ, the above bound says that it is not sufficient to simply match the marginal distributions and achieve small error on the source domain. At the same time, we should also guarantee that the optimal labeling functions (or the conditional distributions of both domains) are not too far away from each other.

It is also helpful to see that $\mathbb{E}_{\mathcal{D}_S} \|f_S - f_T\| = \varepsilon_S(f_T)$ and $\mathbb{E}_{\mathcal{D}_T} [|f_S - f_T|] = \varepsilon_T(f_S)$. In other words, they are essentially the cross-domain errors. When the cross-domain error is small, it implies that the optimal source (resp. target) labeling function generalizes well on the target (resp. source) domain.

Results

One implication of the bound is that when two domains have different marginal label distributions, minimizing the source error while aligning the two domains can lead to increased target error. To verify this, the authors consider the task of digit classification on the MNIST, SVHN and USPS datasets. The label distributions of these three datasets are shown in the figure bellow.

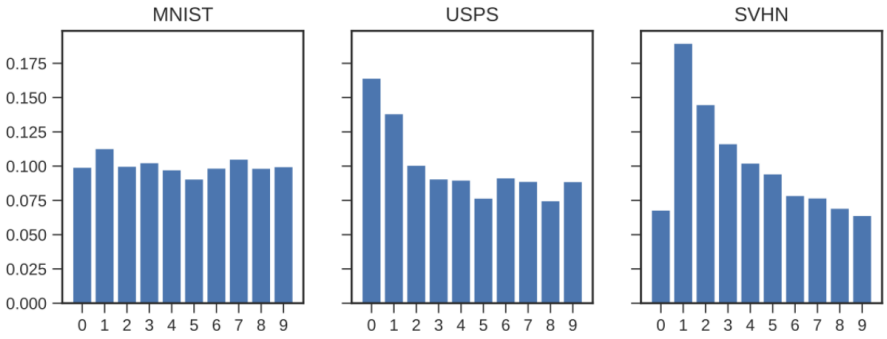

Figure 5: The label (digit) distributions on MNIST, SVHN and USPS.

It is clear to see that these three datasets have quite different label distributions. The figure bellow uses DANN to classify on the target domain by learning a domain invariant representation while training to minimize error on the source domain.

The four adaptation trajectories for DANN are shown in the figure below. Across the four adaptation tasks, we can observe the following pattern: the test domain accuracy rapidly grows within the first 10 iterations before gradually decreasing from its peak, despite consistently increasing source training accuracy. These phase transitions can be verified from the negative slopes of the least squares fit of the adaptation curves (dashed lines). The above experimental results are consistent with the theoretical findings: over-training on the source task can indeed hurt generalization to the target domain when the label distributions differ.

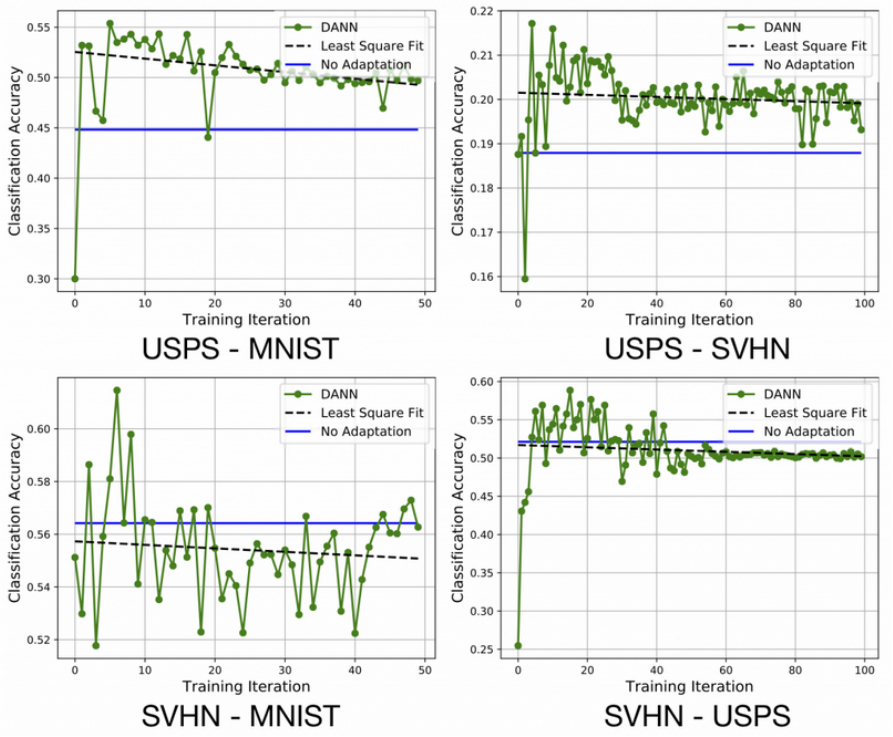

Figure 6: Digit classification on MNIST, USPS and SVHN. The horizontal solid line corresponds to the target domain test accuracy without adaptation. The green solid line is the target domain test accuracy under domain adaptation with DANN. We also plot the least square fit (dashed line) of the DANN adaptation results to emphasize the negative slope.