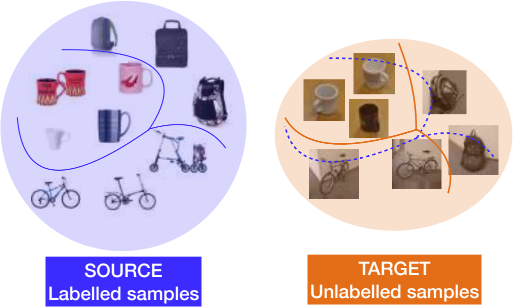
Robust Domain Adaptation: Representations, Weights and Inductive Bias

V. Bouvier, P. Very, C. Chastagnol, M. Tami and C. Hudelot

June 13, 2021

Domain Adaptation



Setup

- **Source labelled samples:** $(x_S^i, y_S^i)_i \sim p_S(X, Y)$
- **Target unlabelled samples:** $(x_T^j)_j \sim p_T(X)$
- **Objective:** Learning $h \in \mathcal{H}$ s.t.: $h \in \arg\min_{h \in \mathcal{H}} \varepsilon_T(h)$

Covariate Shift

Labelling functions are conserved:

$$p_S(y|x) = p_T(y|x) \tag{1}$$

Covariate Shift
Labelling functions are conserved:

$$p_S(y|x) = p_T(y|x) \tag{1}$$

$$\varepsilon_T(h) = \mathbb{E}_T[\ell(h(x), y)] = \mathbb{E}_S\left[\frac{p_T(x)}{p_S(x)}\ell(h(x), y)\right] \tag{2}$$

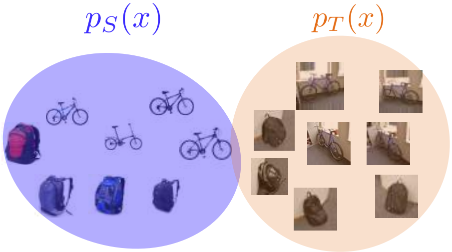
Covariate Shift

Covariate Shift
Labelling functions are conserved:

$$p_S(y|x) = p_T(y|x) \tag{1}$$

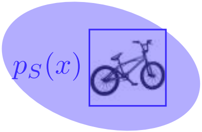
$$\varepsilon_T(h) = \mathbb{E}_T[\ell(h(x), y)] = \mathbb{E}_S\left[\frac{p_T(x)}{p_S(x)}\ell(h(x), y)\right] \tag{2}$$

Needs overlapping supports!

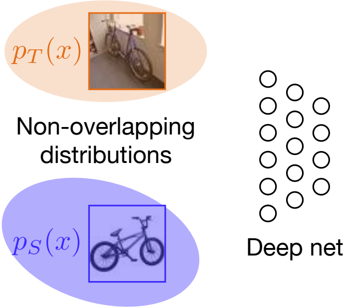


Domain Invariant Representations [Ben-David et al., Ganin et al.]

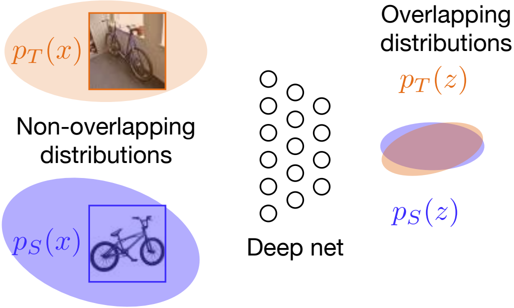
Non-overlapping
distributions



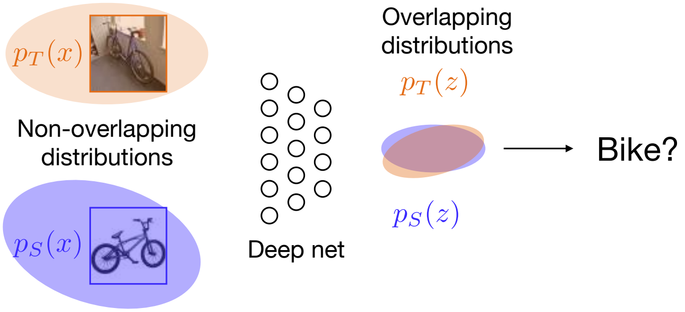
Domain Invariant Representations [Ben-David et al., Ganin et al.]



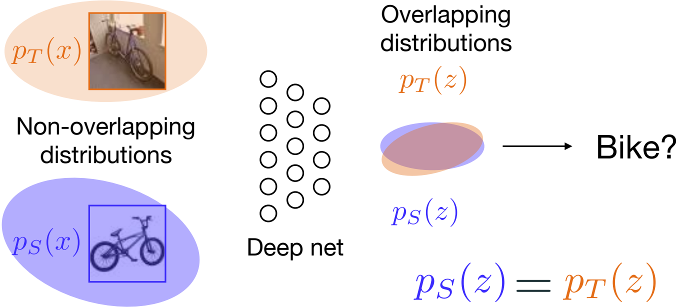
Domain Invariant Representations [Ben-David et al., Ganin et al.]



Domain Invariant Representations [Ben-David et al., Ganin et al.]

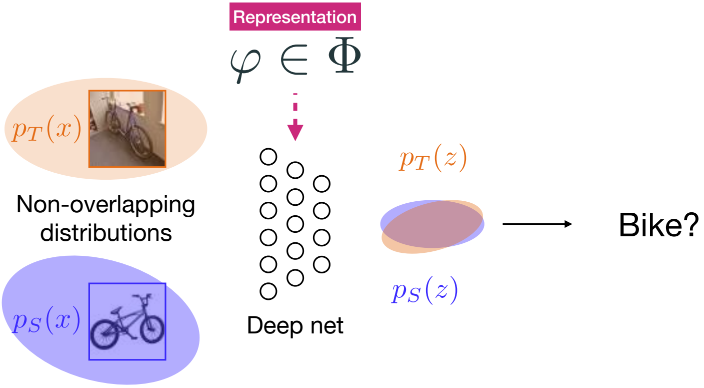


Domain Invariant Representations [Ben-David et al., Ganin et al.]

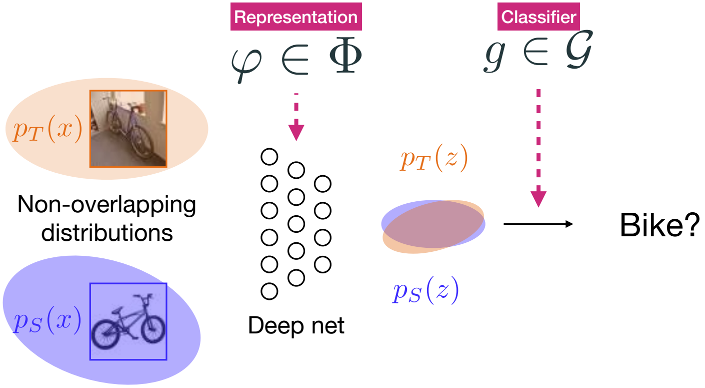


$$p_S(z) = p_T(z)$$

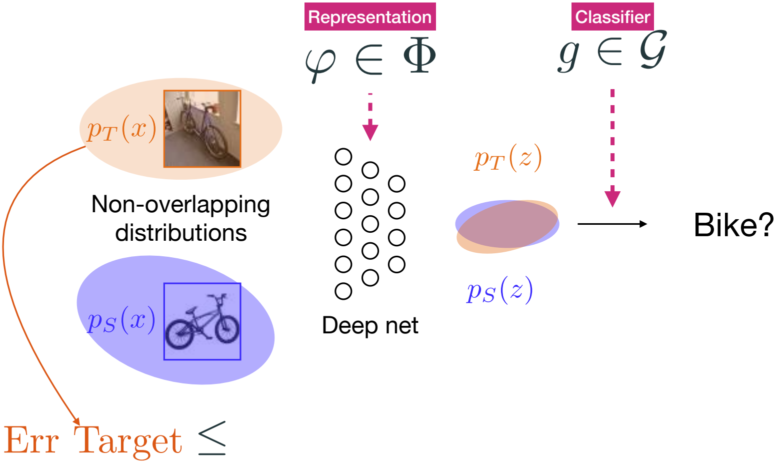
Domain Invariant Representations [Ben-David et al., Ganin et al.]



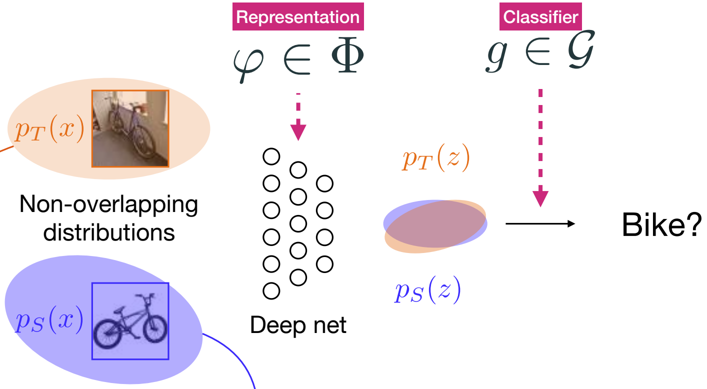
Domain Invariant Representations [Ben-David et al., Ganin et al.]



Domain Invariant Representations [Ben-David et al., Ganin et al.]

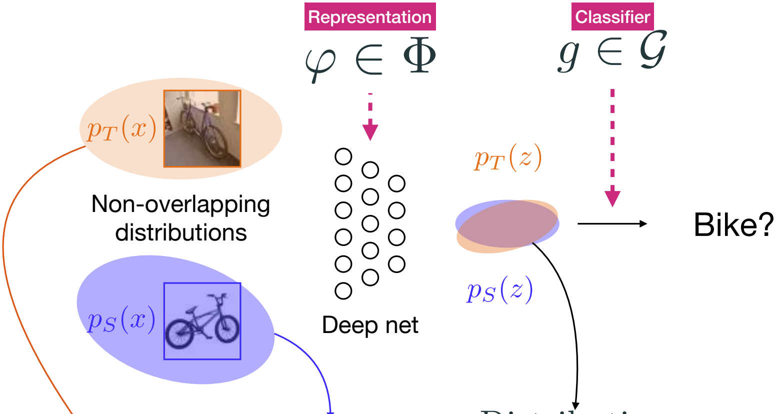


Domain Invariant Representations [Ben-David et al., Ganin et al.]



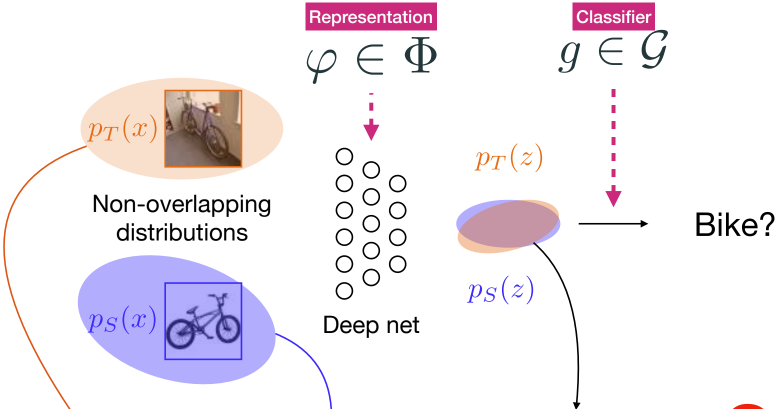
$$\text{Err Target} \leq \text{Err Source}$$

Domain Invariant Representations [Ben-David et al., Ganin et al.]



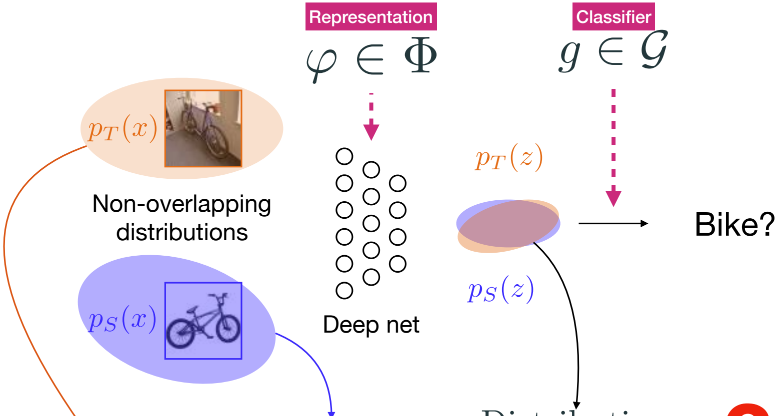
$$\text{Err Target} \leq \text{Err Source} + \begin{array}{c}\text{Distribution}\\\text{divergence}\end{array}$$

Domain Invariant Representations [Ben-David et al., Ganin et al.]



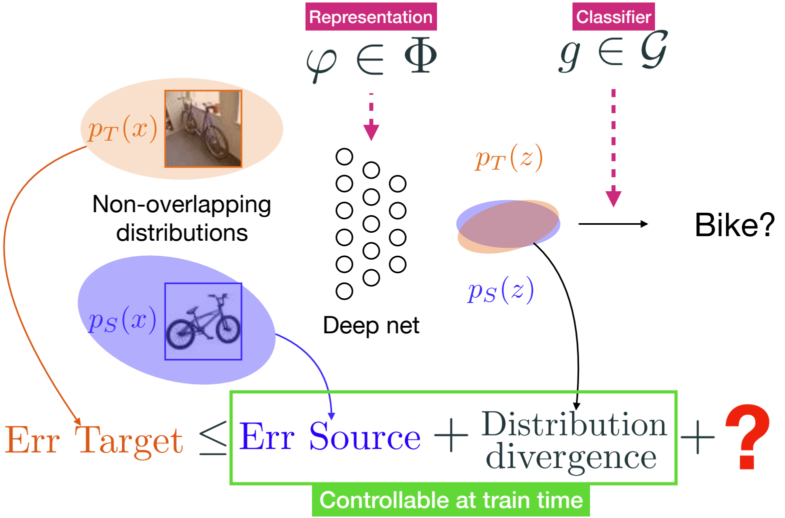
$$\text{Err Target} \leq \text{Err Source} + \text{Distribution divergence} + \text{?}$$

Domain Invariant Representations [Ben-David et al., Ganin et al.]

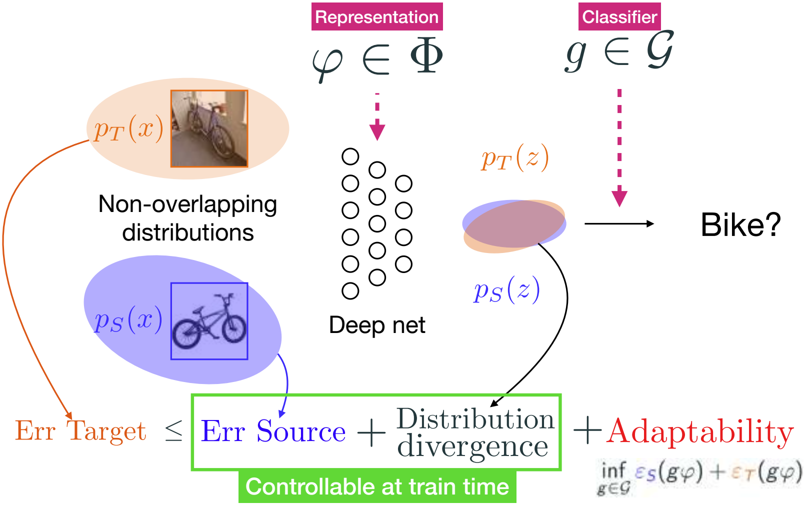


$$\text{Err Target} \leq \text{Err Source} + \begin{array}{c}\text{Distribution}\\ \text{divergence}\end{array} + \textbf{?}$$

Domain Invariant Representations [Ben-David et al., Ganin et al.]



Domain Invariant Representations [Ben-David et al., Ganin et al.]



Domain Invariant Representations: Limits

$$\varepsilon_T(g\varphi) \leq \underbrace{\varepsilon_S(g\varphi) + d_{\mathcal{G}}(\varphi)}_{\text{Controllable}} + \underbrace{\lambda_{\mathcal{G}}(\varphi)}_{\text{Not controllable}} \tag{3}$$

An unexpected trade-off

Let ψ be a representation which is a richer feature extractor than φ:

$$\mathcal{G} \circ \varphi \subset \mathcal{G} \circ \psi$$

Domain Invariant Representations: Limits

$$\varepsilon_T(g\varphi) \leq \underbrace{\varepsilon_S(g\varphi) + d_{\mathcal{G}}(\varphi)}_{\text{Controllable}} + \underbrace{\lambda_{\mathcal{G}}(\varphi)}_{\text{Not controllable}} \tag{3}$$

An unexpected trade-off

Let ψ be a representation which is a richer feature extractor than φ:

$$\mathcal{G} \circ \varphi \subset \mathcal{G} \circ \psi$$

Then,

$$d_{\mathcal{G}}(\varphi) \leq d_{\mathcal{G}}(\psi) \text{ while } \lambda_{\mathcal{G}}(\psi) \leq \lambda_{\mathcal{G}}(\varphi) \tag{4}$$

Domain Invariant Representations: Limits

$$\varepsilon_T(g\varphi) \leq \underbrace{\varepsilon_S(g\varphi) + d_{\mathcal{G}}(\varphi)}_{\text{Controllable}} + \underbrace{\lambda_{\mathcal{G}}(\varphi)}_{\text{Not controllable}} \tag{3}$$

An unexpected trade-off

Let ψ be a representation which is a richer feature extractor than φ:

$$\mathcal{G} \circ \varphi \subset \mathcal{G} \circ \psi$$

Then,

$$d_{\mathcal{G}}(\varphi) \leq d_{\mathcal{G}}(\psi) \text{ while } \lambda_{\mathcal{G}}(\psi) \leq \lambda_{\mathcal{G}}(\varphi) \tag{4}$$

▷ The benefit of representation invariance must be higher than the loss of adaptability, which is impossible to guarantee in practice.

Domain Invariant Representations: Limits

$$\varepsilon_T(g\varphi) \leq \underbrace{\varepsilon_S(g\varphi) + d_{\mathcal{G}}(\varphi)}_{\text{Controllable}} + \underbrace{\lambda_{\mathcal{G}}(\varphi)}_{\text{Not controllable}} \tag{3}$$

An unexpected trade-off

Let ψ be a representation which is a richer feature extractor than φ:

$$\mathcal{G} \circ \varphi \subset \mathcal{G} \circ \psi$$

Then,

$$d_{\mathcal{G}}(\varphi) \leq d_{\mathcal{G}}(\psi) \text{ while } \lambda_{\mathcal{G}}(\psi) \leq \lambda_{\mathcal{G}}(\varphi) \tag{4}$$

▷ The benefit of representation invariance must be higher than the loss of adaptability, which is impossible to guarantee in practice.

Invariance is conflicting with label shift [Zhao et al., ICML2019]

$$\lambda_{\mathcal{G}}(\varphi) \geq \frac{1}{2}\left(\mathrm{JS}(Y) - \mathrm{JS}(Z)\right)^2 \tag{5}$$

If $\mathrm{JS}(Z) \to 0$, $\lambda_{\mathcal{G}}(\varphi)$ can not be small if $\mathrm{JS}(Y)$ is high...

Outline

Outline

1. A new trade-off between invariance and transferability

Outline

1. A new trade-off between invariance and transferability
   - Introduce a new error term named *transferability error*
   - Reconcile Invariant Representations and Weights

Outline

1. A new trade-off between invariance and transferability
   - Introduce a new error term named *transferability error*
   - Reconcile Invariant Representations and Weights
2. Role of **Inductive Bias**:
   - Weights $\triangleright$ *induce new property of invariance of representations and the labelling function*
   - Classifier $\triangleright$ Feedback for better representations invariance.

Outline

1. A new trade-off between invariance and transferability
   - Introduce a new error term named *transferability error*
   - Reconcile Invariant Representations and Weights
2. Role of **Inductive Bias**:
   - Weights ▷ *induce new property of invariance of representations and the labelling function*
   - Classifier ▷ Feedback for better representations invariance.
3. A new algorithm for Robust Unsupervised Domain Adaptation ▷ RUDA
   - Robust to strong label shift
   - Evaluation on two benchmarks

A new trade-off between invariance and transferability

Three ingredients

1. INV $\triangleright$ captures the difference between source and target distribution of representations.

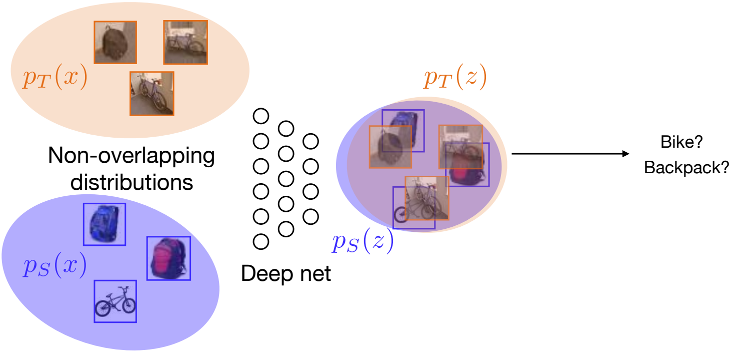
Three ingredients

1. INV ▷ captures the difference between source and target distribution of representations.
2. TSF ▷ catches if the coupling between representations and labels shifts across domains.

Three ingredients

1. INV $\triangleright$ captures the difference between source and target distribution of representations.
2. TSF $\triangleright$ catches if the coupling between representations and labels shifts across domains.
3. Reconcile Weights and Invariant Representations.

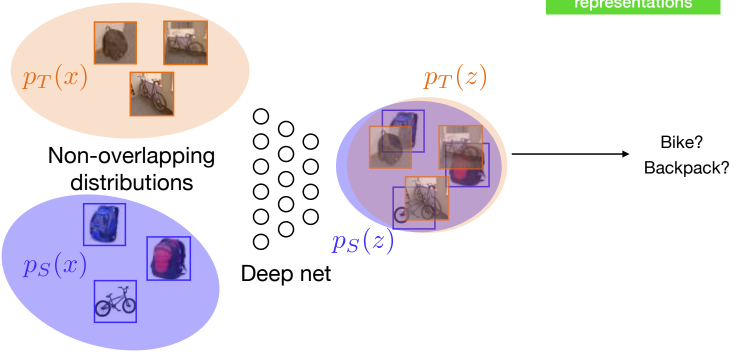
A new trade-off between invariance and transferability



A new trade-off between invariance and transferability

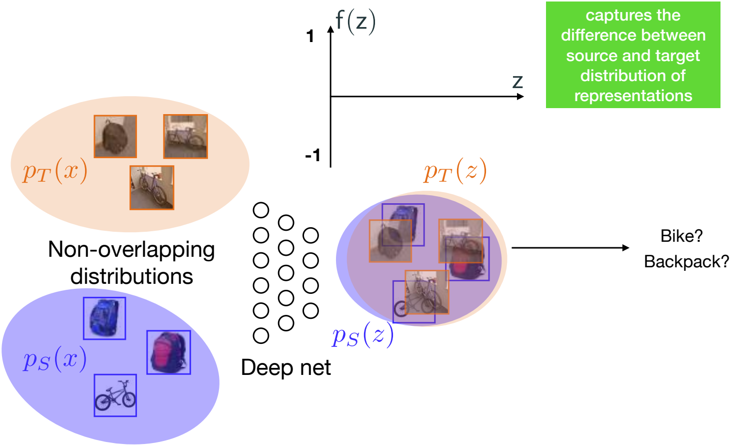
$$\mathrm{INV}(\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[f(Z)] - \mathbb{E}_T[f(Z)]$$

captures the difference between source and target distribution of representations



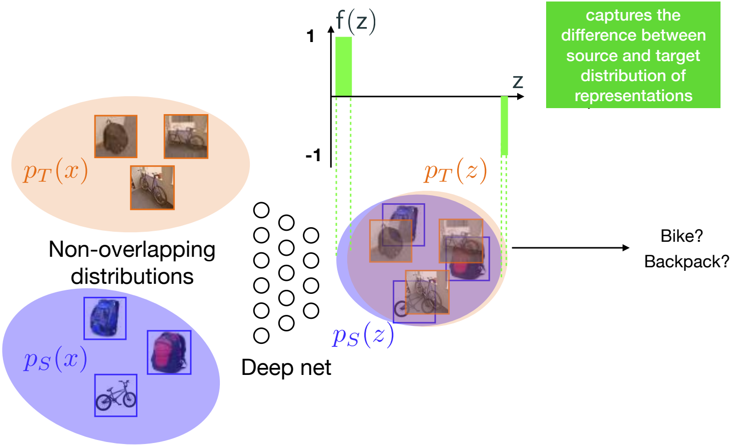
A new trade-off between invariance and transferability

$$\mathrm{INV}(\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[f(Z)] - \mathbb{E}_T[f(Z)]$$



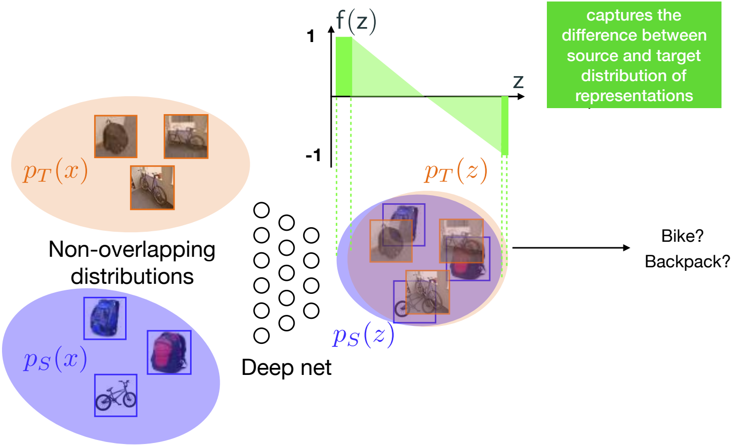
A new trade-off between invariance and transferability

$$\mathrm{INV}(\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[f(Z)] - \mathbb{E}_T[f(Z)]$$



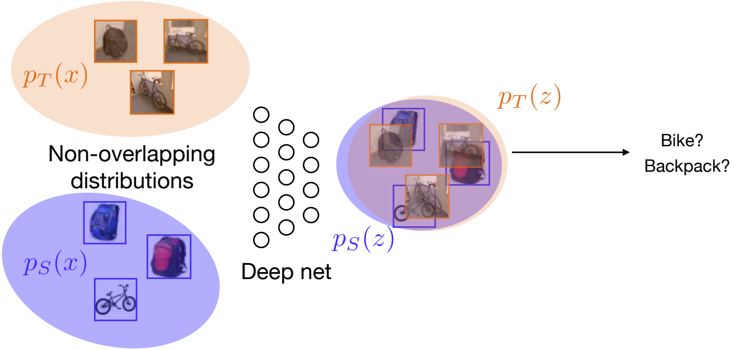
A new trade-off between invariance and transferability

$$\mathrm{INV}(\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[f(Z)] - \mathbb{E}_T[f(Z)]$$



$$\mathrm{TSF}(\varphi) := \sup_{\mathsf{f} \in \mathcal{F}_c} \mathbb{E}_S[Y \cdot \mathsf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathsf{f}(Z)]$$

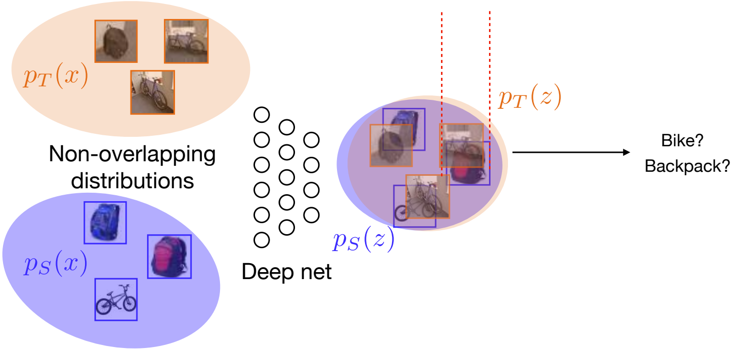
catches if the coupling between representations and labels shifts across domains



A new trade-off between invariance and transferability

$$\mathrm{TSF}(\varphi) := \sup_{\mathsf{f}\in\mathcal{F}_c} \mathbb{E}_S[Y \cdot \mathsf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathsf{f}(Z)]$$

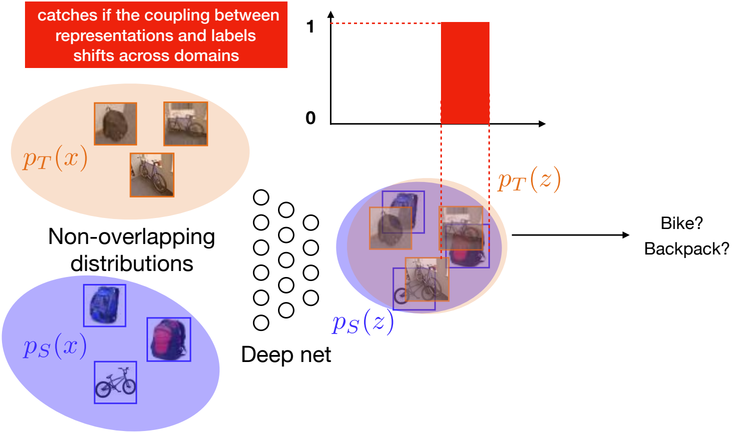
catches if the coupling between representations and labels shifts across domains



A new trade-off between invariance and transferability

$$\mathrm{TSF}(\varphi) := \sup_{f \in \mathcal{F}_c} \mathbb{E}_S[Y \cdot f(Z)] - \mathbb{E}_T[Y \cdot f(Z)]$$

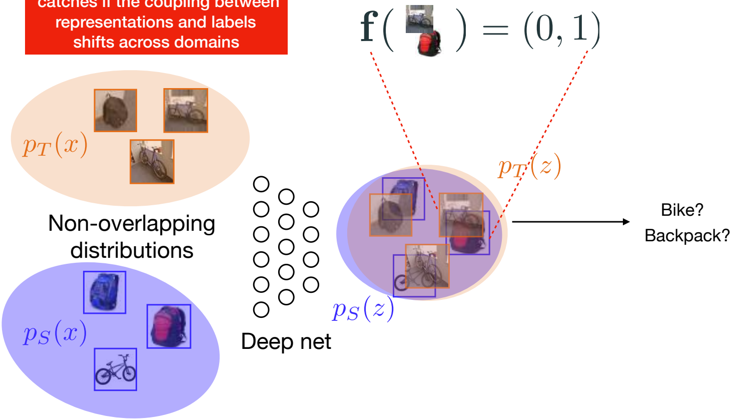
catches if the coupling between representations and labels shifts across domains



A new trade-off between invariance and transferability

$$\mathrm{TSF}(\varphi) := \sup_{\mathsf{f}\in\mathcal{F}_c} \mathbb{E}_S[Y \cdot \mathsf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathsf{f}(Z)]$$

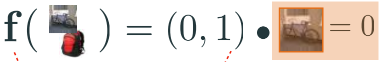
catches if the coupling between representations and labels shifts across domains

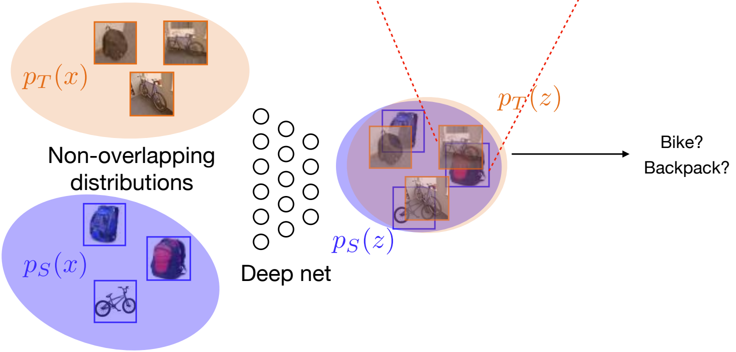


A new trade-off between invariance and transferability

$$\mathrm{TSF}(\varphi) := \sup_{\mathsf{f} \in \mathcal{F}_c} \mathbb{E}_S[Y \cdot \mathsf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathsf{f}(Z)]$$

catches if the coupling between representations and labels shifts across domains

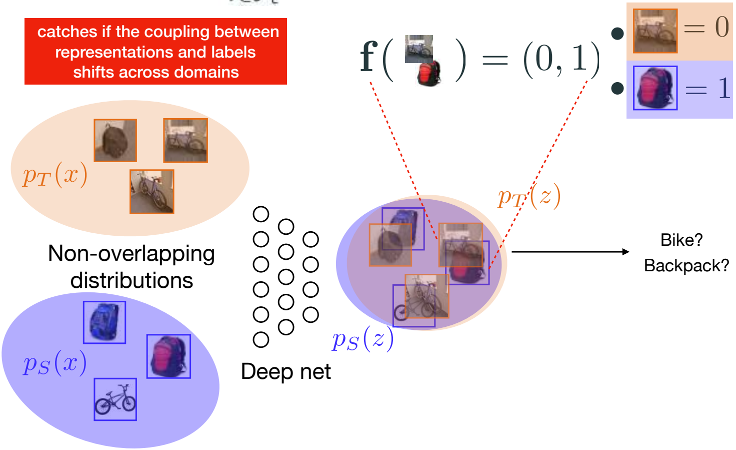




A new trade-off between invariance and transferability

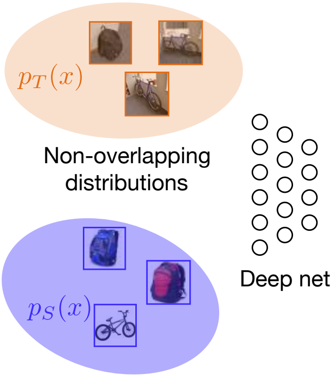
$$\mathrm{TSF}(\varphi) := \sup_{f \in \mathcal{F}_c} \mathbb{E}_S[Y \cdot f(Z)] - \mathbb{E}_T[Y \cdot f(Z)]$$

catches if the coupling between representations and labels shifts across domains

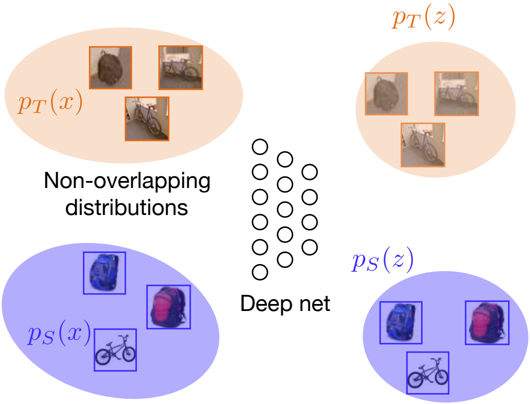


Reconciling Weights and Invariant Representations

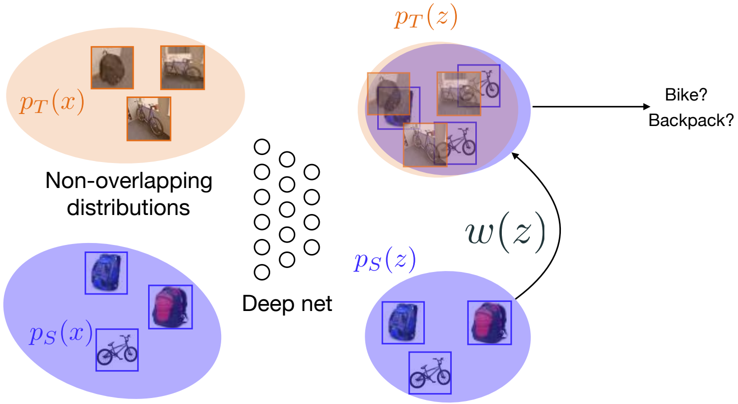
Reconciling Weights and Invariant Representations



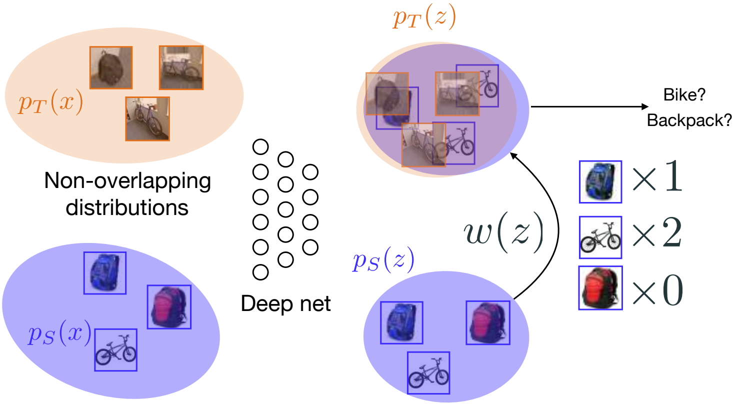
Reconciling Weights and Invariant Representations



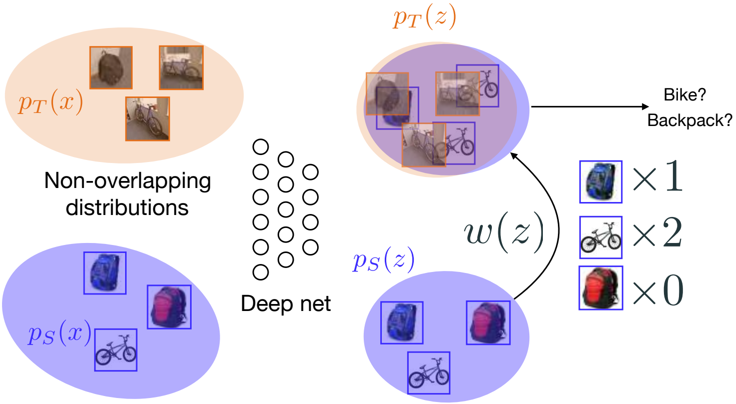
Reconciling Weights and Invariant Representations



Reconciling Weights and Invariant Representations

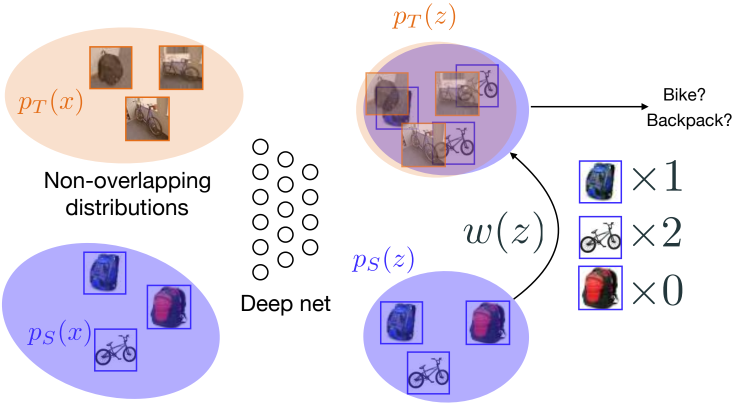


Reconciling Weights and Invariant Representations



$$\mathrm{INV}(w, \varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[w(Z)f(Z)] - \mathbb{E}_T[f(Z)]$$

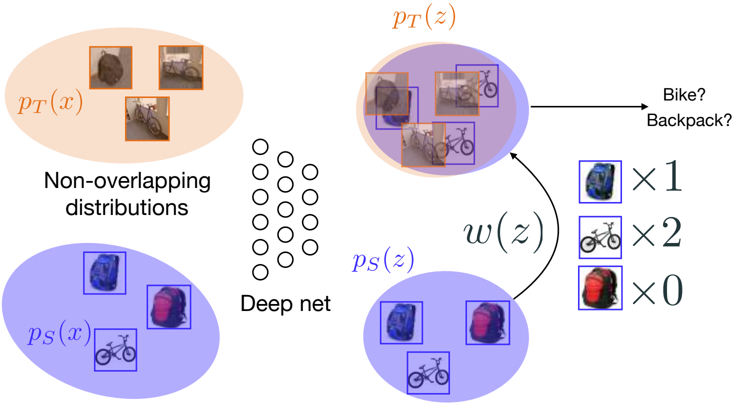
Reconciling Weights and Invariant Representations



$$\mathrm{INV}(w,\varphi) := \sup_{f\in\mathcal{F}} \mathbb{E}_S[w(Z)f(Z)] - \mathbb{E}_T[f(Z)]$$

$$\mathrm{TSF}(w,\varphi) := \sup_{\mathbf{f}\in\mathcal{F}_c} \mathbb{E}_S[w(Z)Y\cdot \mathbf{f}(Z)] - \mathbb{E}_T[Y\cdot \mathbf{f}(Z)]$$

Reconciling Weights and Invariant Representations

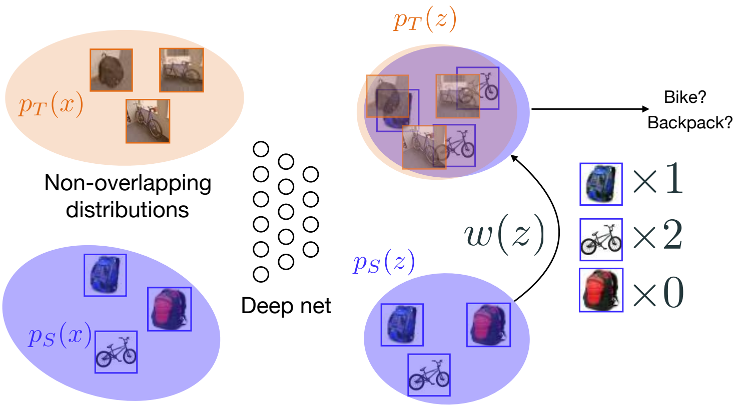


$$\mathrm{INV}(w, \varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[w(Z)f(Z)] - \mathbb{E}_T[f(Z)]$$

$$\mathrm{TSF}(w, \varphi) := \sup_{\mathbf{f} \in \mathcal{F}_c} \mathbb{E}_S[w(Z)Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathbf{f}(Z)]$$

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi)$$

Reconciling Weights and Invariant Representations

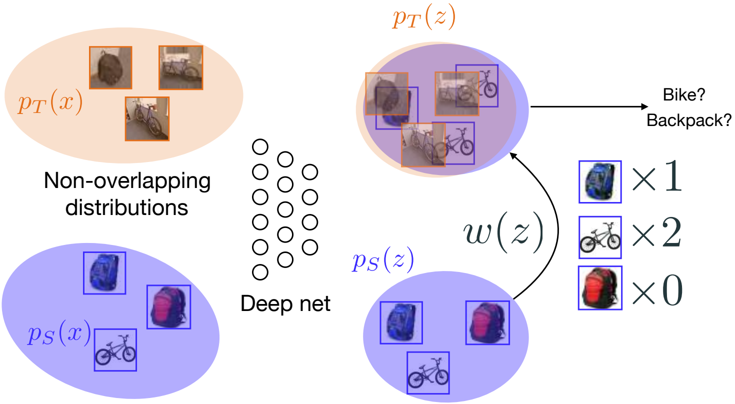


$$\mathrm{INV}(w, \varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[w(Z) f(Z)] - \mathbb{E}_T[f(Z)]$$

$$\mathrm{TSF}(w, \varphi) := \sup_{\mathsf{f} \in \mathcal{F}_c} \mathbb{E}_S[w(Z) Y \cdot \mathsf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathsf{f}(Z)]$$

$$\boxed{\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w, \varphi)}$$

Reconciling Weights and Invariant Representations

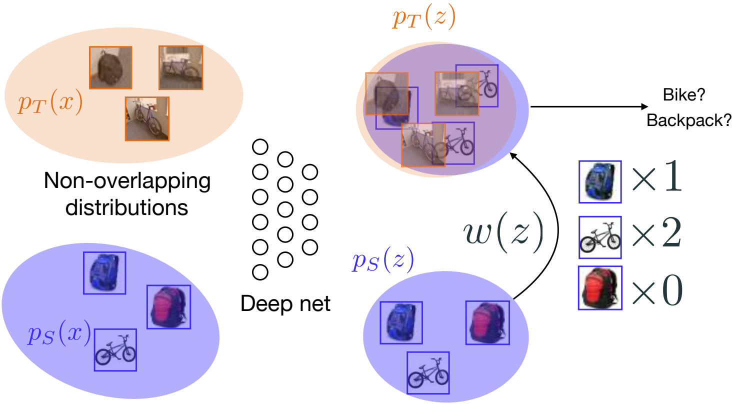


$$\mathrm{INV}(w,\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[w(Z)f(Z)] - \mathbb{E}_T[f(Z)]$$

$$\mathrm{TSF}(w,\varphi) := \sup_{\mathbf{f} \in \mathcal{F}_c} \mathbb{E}_S[w(Z)Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathbf{f}(Z)]$$

$$\boxed{\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w,\varphi) + 2 \cdot \mathrm{TSF}(w,\varphi)}$$

Reconciling Weights and Invariant Representations



$$\mathrm{INV}(w,\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[w(Z)f(Z)] - \mathbb{E}_T[f(Z)]$$

$$\mathrm{TSF}(w,\varphi) := \sup_{\mathbf{f} \in \mathcal{F}_c} \mathbb{E}_S[w(Z)Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathbf{f}(Z)]$$

$$\boxed{\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w,\varphi) + 2 \cdot \mathrm{TSF}(w,\varphi) + \varepsilon_T(\mathbf{f}_T\varphi)}$$

Reconciling Weights and Invariant Representations

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w,\varphi) + 2 \cdot \mathrm{TSF}(w,\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (6)$$

Reconciling Weights and Invariant Representations

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w\cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w, \varphi) + 2 \cdot \mathrm{TSF}(w, \varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (6)$$

Weighting the source domain

$$\varepsilon_{w\cdot S}(g\varphi) := \mathbb{E}_S[w(Z)\ell(g(Z), Y)] \quad (7)$$

A new trade-off

$$\varepsilon_T(\mathbf{f}_T\varphi) := \inf_{\mathbf{f} \in \mathcal{F}_c} \varepsilon_T(\mathbf{f}\varphi) \quad (8)$$

Reconciling Weights and Invariant Representations

$$\varepsilon_T(g\varphi) \le \varepsilon_{w\cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w,\varphi) + 2 \cdot \mathrm{TSF}(w,\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (6)$$

Weighting the source domain

$$\varepsilon_{w\cdot S}(g\varphi) := \mathbb{E}_S[w(Z)\ell(g(Z),Y)] \quad (7)$$

A new trade-off

$$\varepsilon_T(\mathbf{f}_T\varphi) := \inf_{\mathbf{f}\in\mathcal{F}_c} \varepsilon_T(\mathbf{f}\varphi) \quad (8)$$

Tightness

$\mathrm{INV}(w,\varphi) = 0$ and $\mathrm{TSF}(w,\varphi) = 0$, then,

$$\implies p_T(y|z) = p_S(y|z) \quad \text{and} \quad w(z) = \frac{p_T(z)}{p_S(z)}$$

▷ Much smaller than the adaptability.

Remaining challenges

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w\cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w,\varphi) + 2 \cdot \mathrm{TSF}(w,\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (9)$$

Remaining challenges

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w, \varphi) + 2 \cdot \mathrm{TSF}(w, \varphi) + \varepsilon_T(\mathbf{f}_T \varphi) \quad (9)$$

1. Classifier ▷ *address the lack of labelled data in the target domain.*

$$\mathrm{TSF}(w, \varphi) := \sup_{\mathbf{f} \in \mathcal{F}_c} \mathbb{E}_S[w(Z)Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\underbrace{Y}_{??} \cdot \mathbf{f}(Z)]$$

2. Weights ▷ induce invariance property on representations (see the paper for more details).

Remaining challenges

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w\cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w,\varphi) + 2 \cdot \mathrm{TSF}(w,\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (9)$$

1. $\boxed{\textit{Classifier} \triangleright \textit{address the lack of labelled data in the target domain.}}$

$$\mathrm{TSF}(w,\varphi) := \sup_{\mathbf{f}\in\mathcal{F}_c} \mathbb{E}_S[w(Z)Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\underbrace{Y}_{??} \cdot \mathbf{f}(Z)]$$

2. Weights $\triangleright$ induce invariance property on representations (see the paper for more details).

Inductive design of the classifier

Inductive design of the classifier

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f} \in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \tag{10}$$

Inductive design of the classifier

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f} \in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \tag{10}$$

Insight

The best source classifier is not the best target classifier, and, it is possible to improve the best source classifier , *e.g.*, specific architecture or a well-suited regularization.

Inductive design of the classifier

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f} \in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \tag{10}$$

Insight

The best source classifier is not the best target classifier, and, it is possible to improve the best source classifier , *e.g.*, specific architecture or a well-suited regularization.

Inductive design

We say that there is an inductive design of a classifier at level $0 < \beta \leq 1$ if for any representations φ, noting $g_S = \arg\min_{g \in \mathcal{G}} \varepsilon_S(g\varphi)$, we can determine $\tilde{g}$ such that:

$$\varepsilon_T(\tilde{g}\varphi) \leq \beta \varepsilon_T(g_S \varphi) \tag{11}$$

- $\beta-$strong when $\beta < 1$,
- weak when $\beta = 1$.

Generalization guarantees with inductive bias

A revisited version of the bound:

$$\varepsilon_T(g_S\varphi) \leq \varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\tilde{g}\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (12)$$

Generalization guarantees with inductive bias

A revisited version of the bound:

$$\varepsilon_T(g_S\varphi) \leq \varepsilon_S(g_S\varphi) + 6\cdot\mathrm{INV}(\varphi) + 2\cdot\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\tilde{g}\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (12)$$

- $\widehat{\mathrm{TSF}}(\varphi, \tilde{g})$ ▷ *transferability of representations with respect to the inductive design:*

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f}\in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \quad (13)$$

Generalization guarantees with inductive bias

A revisited version of the bound:

$$\varepsilon_T(g_S\varphi) \leq \varepsilon_S(g_S\varphi) + 6\cdot\mathrm{INV}(\varphi) + 2\cdot\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\tilde{g}\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (12)$$

- $\widehat{\mathrm{TSF}}(\varphi, \tilde{g})$ ▷ *transferability of representations with respect to the inductive design:*

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f}\in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \quad (13)$$

- $\varepsilon_T(\tilde{g}\varphi)$ ▷ *How good is the inductive design.*

Generalization guarantees with inductive bias

A revisited version of the bound:

$$\varepsilon_T(g_S\varphi) \leq \varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\tilde{g}\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (12)$$

- $\widehat{\mathrm{TSF}}(\varphi, \tilde{g})$ ▷ *transferability of representations with respect to the inductive design:*

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f} \in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \quad (13)$$

- $\varepsilon_T(\tilde{g}\varphi)$ ▷ *How good is the inductive design.*

Here comes the guarantees

$$\boxed{\varepsilon_T(g_S\varphi)} \leq \varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \beta \boxed{\varepsilon_T(g_S\varphi)} + \varepsilon_T(\mathbf{f}_T\varphi)$$

Generalization guarantees with inductive bias

A revisited version of the bound:

$$\varepsilon_T(g_S\varphi) \leq \varepsilon_S(g_S\varphi)+6\cdot\mathrm{INV}(\varphi)+2\cdot\widehat{\mathrm{TSF}}(\varphi,\tilde{g})+\varepsilon_T(\tilde{g}\varphi)+\varepsilon_T(\mathbf{f}_T\varphi) \quad (12)$$

- $\widehat{\mathrm{TSF}}(\varphi,\tilde{g})$ ▷ *transferability of representations with respect to the inductive design:*

$$\widehat{\mathrm{TSF}}(\varphi,\tilde{g}) := \sup_{\mathbf{f}\in F_C} \mathbb{E}_S[Y\cdot\mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z)\cdot\mathbf{f}(Z)] \quad (13)$$

- $\varepsilon_T(\tilde{g}\varphi)$ ▷ *How good is the inductive design.*

Here comes the guarantees

$$\boxed{\varepsilon_T(g_S\varphi)} \leq \frac{1}{1-\beta}\Big(\varepsilon_S(g_S\varphi)+6\cdot\mathrm{INV}(\varphi)+2\cdot\widehat{\mathrm{TSF}}(\varphi,\tilde{g})+\varepsilon_T(\mathbf{f}_T\varphi)\Big)$$

Generalization guarantees with inductive bias

A revisited version of the bound:

$$\varepsilon_T(g_S\varphi) \leq \varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\tilde{g}\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \quad (12)$$

- $\widehat{\mathrm{TSF}}(\varphi, \tilde{g})$ ▷ *transferability of representations with respect to the inductive design:*

$$\widehat{\mathrm{TSF}}(\varphi, \tilde{g}) := \sup_{\mathbf{f} \in F_C} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[\tilde{g}(Z) \cdot \mathbf{f}(Z)] \quad (13)$$

- $\varepsilon_T(\tilde{g}\varphi)$ ▷ *How good is the inductive design.*

Here comes the guarantees

$$\varepsilon_T(\tilde{g}\varphi) \leq \frac{\beta}{1-\beta}\Big(\varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\mathbf{f}_T\varphi)\Big)$$

Generalization guarantees with inductive bias

$$\varepsilon_T(\tilde{g}\varphi) \leq \frac{\beta}{1-\beta}\Big(\varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\mathbf{f}_T\varphi)\Big)$$

Generalization guarantees with inductive bias

$$\varepsilon_T(\tilde{g}\varphi) \leq \frac{\beta}{1-\beta}\left(\varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\mathbf{f}_T\varphi)\right)$$

Generalization guarantees with inductive bias

$$\varepsilon_T(\tilde{\mathbf{g}}\varphi) \leq \frac{\beta}{1-\beta}\Big(\varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{\mathbf{g}}) + \varepsilon_T(\mathbf{f}_T\varphi)\Big)$$

- Target labels are only involved in $\varepsilon_T(\mathbf{f}_T\varphi)$ which reflects the level of noise when fitting labels from representations ▷ *transferability is now free of target labels.*

Generalization guarantees with inductive bias

$$\varepsilon_T(\tilde{\mathbf{g}}\varphi) \leq \frac{\beta}{1-\beta}\left(\varepsilon_S(g_S\varphi) + 6\cdot \mathrm{INV}(\varphi) + 2\cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{\mathbf{g}}) + \varepsilon_T(\mathbf{f}_T\varphi)\right)$$

- Target labels are only involved in $\varepsilon_T(\mathbf{f}_T\varphi)$ which reflects the level of noise when fitting labels from representations $\triangleright$ *transferability is now free of target labels.*
- the weaker the inductive bias ($\beta \to 1$), the higher the bound and vice versa.

Generalization guarantees with inductive bias

$$\varepsilon_T(\tilde{g}\varphi) \leq \frac{\beta}{1-\beta}\Big(\varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{g}) + \varepsilon_T(\mathbf{f}_T\varphi)\Big)$$

- Target labels are only involved in $\varepsilon_T(\mathbf{f}_T\varphi)$ which reflects the level of noise when fitting labels from representations ▷ *transferability is now free of target labels.*
- the weaker the inductive bias ($\beta \to 1$), the higher the bound and vice versa.
- **Takeaways** ▷ if a regularization is available, it will interacts with the transferability error.

Robust Unsupervised Domain Adaptation (RUDA)

Assumptions

- Weak inductive design of the classifier:
  - **Classifier**: $\tilde{g} \leftarrow g_S$ $(\beta = 1)$
  - *No theoretical guarantees from inductive classifier.*

Assumptions

- Weak inductive design of the classifier:
  - **Classifier**: $\tilde{g} \leftarrow g_S$ $(\beta = 1)$
  - *No theoretical guarantees from inductive classifier.*
- Weights controls the invariance error: $w(z) = \frac{p_T(z)}{p_S(z)}$

Assumptions

- Weak inductive design of the classifier:
  - **Classifier**: $\tilde{g} \leftarrow g_S$ $(\beta = 1)$
  - *No theoretical guarantees from inductive classifier.*
- Weights controls the invariance error: $w(z) = \frac{p_T(z)}{p_S(z)}$
- Bring strong robustness to the adaptation procedure $\triangleright$ *stress-test on dataset with label strong label shift.*

Assumptions

- Weak inductive design of the classifier:
  - **Classifier**: $\tilde{g} \leftarrow g_S$ $(\beta = 1)$
  - *No theoretical guarantees from inductive classifier.*
- Weights controls the invariance error: $w(z) = \frac{p_T(z)}{p_S(z)}$
- Bring strong robustness to the adaptation procedure ▷ *stress-test on dataset with label strong label shift.*

$$\begin{cases} \varphi^\star & = \arg\min_{\varphi \in \Phi} \ \varepsilon_{w(\varphi)\cdot S}(g_{w\cdot S}\varphi) + \lambda \cdot \widehat{\mathrm{TSF}}(w, \varphi, g_{w\cdot S}) \\ & \text{such that } w(\varphi) = \arg\min_{w} \mathrm{INV}(w, \varphi) \end{cases}$$

▷ More details about the procedure in the paper.

Experiments

Table 1: Accuracy (%) on the **Office-31** dataset.

	Method	A→W	W→A	A→D	D→A	D→W	W→D	Avg
Standard	ResNet-50	68.4 ± 0.2	60.7 ± 0.3	68.9 ± 0.2	62.5 ± 0.3	96.7 ± 0.1	99.3 ± 0.1	76.1
	DANN	82.0 ± 0.4	67.4 ± 0.5	79.7 ± 0.4	68.2 ± 0.4	96.9 ±0.2	99.1 ±0.1	82.2
	CDAN	93.1 ± 0.2	68.0 ± 0.4	89.8 ± 0.3	70.1 ± 0.4	98.2 ± 0.2	100. ± 0.0	86.6
	CDAN+E	94.1 ± 0.1	69.3 ± 0.4	**92.9 ± 0.2**	**71.0 ± 0.3**	**98.6 ± 0.1**	**100. ± 0.0**	**87.7**
	RUDA	**94.3 ± 0.3**	**70.7 ± 0.3**	92.1 ± 0.3	70.7 ± 0.1	98.5 ± 0.1	100. ± 0.0	87.6
	$RUDA_w$	92.0 ± 0.3	67.9 ± 0.3	91.1 ± 0.3	70.2 ± 0.2	98.6 ± 0.1	100. ± 0.0	86.6
$5 \times [16 \sim 31]$	ResNet-50	72.4 ± 0.7	59.5 ± 0.1	79.0 ± 0.1	61.6 ± 0.3	97.8 ± 0.1	99.3 ± 0.1	78.3
	DANN	67.5 ± 0.1	52.1 ± 0.8	69.7 ± 0.0	51.5 ± 0.1	89.9 ±0.1	75.9 ±0.2	67.8
	CDAN	82.5 ± 0.4	62.9 ± 0.6	81.[illegible] ± 0.5	65.5 ± 0.5	98.5 ± 0.3	99.8 ± 0.0	81.6
	RUDA	85.4 ± 0.8	66.7 ± 0.5	81.3 ± 0.3	64.0 ± 0.5	98.4 ± 0.2	99.5 ± 0.1	82.1
	IWAN	72.4 ± 0.4	54.8 ± 0.8	75.0 ± 0.3	54.8 ± 1.3	97.0 ±0.0	95.8 ±0.6	75.0
	$CDAN_w$	81.5 ±0.5	64.5 ± 0.4	80.7 ± 1.0	65 ± 0.8	**98.7 ± 0.2**	99.9 ± 0.1	81.8
	$RUDA_w$	**87.4 ± 0.2**	**68.3 ± 0.3**	**82.9** ± 0.4	**68.8 ± 0.2**	**98.7** ± 0.1	**100.** ± 0.0	**83.8**

RUDA performs similarly than SOTA approaches

Table 2: Accuracy (%) on the **Digits** dataset.

Method	U→ M						M→U						
Shift of [0 ∼ 5]	5%	10%	15%	20%	100%	Avg	5%	10%	15%	20%	100%	Avg	Avg
DANN	41.7	51.0	59.6	69.0	94.5	63.2	34.5	51.0	59.6	63.6	90.7	59.9	63.2
CDAN	50.7	62.2	82.9	82.8	**96.9**	75.1	32.0	69.7	78.9	81.3	**93.9**	71.2	73.2
RUDA	44.4	58.4	80.0	84.0	95.5	72.5	34.9	59.0	76.1	78.8	93.3	68.4	70.5
IWAN	73.7	74.4	78.4	77.5	95.7	79.9	72.2	82.0	84.3	86.0	92.0	83.3	81.6
$CDAN_w$	68.3	78.8	84.9	**88.4**	96.6	83.4	69.4	80.0	83.5	87.8	93.7	82.9	83.2
$RUDA_w$	**78.7**	**82.8**	**86.0**	86.9	93.9	**85.7**	**78.7**	**87.9**	**88.2**	**89.3**	92.5	**87.3**	**86.5**

Experiments

Table 1: Accuracy (%) on the **Office-31** dataset.

	Method	A→W	W→A	A→D	D→A	D→W	W→D	Avg
Standard	ResNet-50	68.4 ± 0.2	60.7 ± 0.3	68.9 ± 0.2	62.5 ± 0.3	96.7 ± 0.1	99.3 ± 0.1	76.1
	DANN	82.0 ± 0.4	67.4 ± 0.5	79.7 ± 0.4	68.2 ± 0.4	96.9 ±0.2	99.1 ±0.1	82.2
	CDAN	93.1 ± 0.2	68.0 ± 0.4	89.8 ± 0.3	70.1 ± 0.4	98.2 ± 0.2	100. ± 0.0	86.6
	CDAN+E	94.1 ± 0.1	69.3 ± 0.4	**92.9 ± 0.2**	**71.0 ± 0.3**	**98.6 ± 0.1**	**100. ± 0.0**	**87.7**
	RUDA	**94.3 ± 0.3**	**70.7 ± 0.3**	92.1 ± 0.3	70.7 ± 0.1	98.5 ± 0.1	100. ± 0.0	87.6
	$RUDA_w$	92.0 ± 0.3	67.9 ± 0.3	91.1 ± 0.3	70.2 ± 0.2	98.6 ± 0.1	100. ± 0.0	86.6
$5 \times [16 \sim 31]$	ResNet-50	72.4 ± 0.7	59.5 ± 0.1	79.0 ± 0.1	61.6 ± 0.3	97.8 ± 0.1	99.3 ± 0.1	78.3
	DANN	67.5 ± 0.1	52.1 ± 0.8	69.7 ± 0.0	51.5 ± 0.1	89.9 ±0.1	75.9 ±0.2	67.8
	CDAN	82.5 ± 0.4	62.9 ± 0.6	81.4 ± 0.5	65.5 ± 0.5	98.5 ± 0.3	99.8 ± 0.0	81.6
	RUDA	85.4 ± 0.8	66.7 ± 0.5	81.3 ± 0.3	64.0 ± 0.5	98.4 ± 0.2	99.5 ± 0.1	82.1
	IWAN	72.4 ± 0.4	54.8 ± 0.8	75.0 ± 0.3	54.8 ± 1.3	97.0 ±0.0	95.8 ±0.6	75.0
	$CDAN_w$	81.5 ±0.5	64.5 ± 0.4	80.7 ± 1.0	65 ± 0.8	**98.7 ± 0.2**	99.9 ± 0.1	81.8
	$RUDA_w$	**87.4 ± 0.2**	**68.3 ± 0.3**	**82.9** ± 0.4	**68.8 ± 0.2**	**98.7** ± 0.1	**100.** ± 0.0	**83.8**

RUDA still performs well even when stress with strong label shift

Table 2: Accuracy (%) on the **Digits** dataset.

Method / Shift of $[0 \sim 5]$	U→M 5%	10%	15%	20%	100%	Avg	M→U 5%	10%	15%	20%	100%	Avg	Avg
DANN	41.7	51.0	59.6	69.0	94.5	63.2	34.5	51.0	59.6	63.6	90.7	59.9	63.2
CDAN	50.7	62.2	82.9	82.8	**96.9**	75.1	32.0	69.7	78.9	81.3	**93.9**	71.2	73.2
RUDA	44.4	58.4	80.0	84.0	95.5	72.5	34.9	59.0	76.1	78.8	93.3	68.4	70.5
IWAN	73.7	74.4	78.4	77.5	95.7	79.9	72.2	82.0	84.3	86.0	92.0	83.3	81.6
$CDAN_w$	68.3	78.8	84.9	**88.4**	96.6	83.4	69.4	80.0	83.5	87.8	93.7	82.9	83.2
$RUDA_w$	**78.7**	**82.8**	**86.0**	86.9	93.9	**85.7**	**78.7**	**87.9**	**88.2**	**89.3**	92.5	**87.3**	**86.5**

Conclusion

Conclusion

1. New bound of the target risk which unifies weights and representations in UDA.

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w, \varphi) + 2 \cdot \mathrm{TSF}(w, \varphi) + \varepsilon_T(\mathbf{f}_T\varphi)$$

Conclusion

1. New bound of the target risk which unifies weights and representations in UDA.
2. Theoretical analysis of the role of inductive bias when designing both weights and the classifier.

$$\varepsilon_T(\tilde{\mathbf{g}}\varphi) \leq \frac{\beta}{1-\beta}\left(\varepsilon_S(g_S\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \widehat{\mathrm{TSF}}(\varphi, \tilde{\mathbf{g}}) + \varepsilon_T(\mathbf{f}_T\varphi)\right)$$

Conclusion

1. New bound of the target risk which unifies weights and representations in UDA.
2. Theoretical analysis of the role of inductive bias when designing both weights and the classifier.
3. New learning procedure $\triangleright$ *weak inductive bias can make adaptation more robust even when stressed by strong label shift between source and target domains.*

$$\begin{cases} \varphi^\star & = \arg\min_{\varphi \in \Phi} \ \varepsilon_{w(\varphi)\cdot S}(g_{w\cdot S}\varphi) + \lambda \cdot \widehat{\mathrm{TSF}}(w, \varphi, g_{w\cdot S}) \\ & \text{such that } w(\varphi) = \arg\min_{w} \mathrm{INV}(w, \varphi) \end{cases}$$

Conclusion

1. New bound of the target risk which unifies weights and representations in UDA.
2. Theoretical analysis of the role of inductive bias when designing both weights and the classifier.
3. New learning procedure ▷ *weak inductive bias can make adaptation more robust even when stressed by strong label shift between source and target domains.*

This work leaves room for in-depth study of stronger inductive bias by providing both theoretical and empirical foundations.

Thank you!

Motivations

$$\varepsilon_T(g\varphi) \leq \underbrace{\varepsilon_S(g\varphi) + d_{\mathcal{G}}(\varphi)}_{\text{Controllable}} + \underbrace{\lambda_{\mathcal{G}}(\varphi)}_{\text{Not controllable}} \tag{14}$$

An unexpected trade-off

Let ψ be a representation which is a richer feature extractor than φ: $\mathcal{G} \circ \varphi \subset \mathcal{G} \circ \psi$. Then,

$$d_{\mathcal{G}}(\varphi) \leq d_{\mathcal{G}}(\psi) \text{ while } \lambda_{\mathcal{G}}(\psi) \leq \lambda_{\mathcal{G}}(\varphi) \tag{15}$$

Motivations

Invariance is conflicting with label shift [Zhao et al., ICML2019]

$$\lambda_{\mathcal{G}}(\varphi) \geq \frac{1}{2}\left(\mathrm{JS}(Y) - \mathrm{JS}(Z)\right)^2 \tag{16}$$

If $\mathrm{JS}(Z) \to 0$, $\lambda_{\mathcal{G}}(\varphi)$ can not be small if $\mathrm{JS}(Y)$ is high... $\triangleright$ *We should weight the source distributions! But how...*

- how to design weights?
- how weights interact with invariance?
- why predictions are important in UDA?

Our strategy

▷ emerges from the sup / inf duality computed on a small hypothesis class $\mathcal{G} \circ \varphi$.

Our strategy

Express both invariance and transferability of representations as supremum over a large space of critic functions

▷ See the paper for details about property of the critic functions

2 critic functions space

- $\mathcal{F}$ ▷ 'large' function space from $\mathcal{Z}$ to $[-1, 1]$
- $\mathcal{F}_C$ ▷ 'large' function space from $\mathcal{Z}$ to $[-1, 1]^C$

▷ Typically continuous functions.

Tools

2 critic functions space

- $\mathcal{F}$ ▷ 'large' function space from $\mathcal{Z}$ to $[-1, 1]$
- $\mathcal{F}_C$ ▷ 'large' function space from $\mathcal{Z}$ to $[-1, 1]^C$

▷ Typically continuous functions.

2 errors

- captures the difference between source and target distribution of representations:

$$\mathrm{INV}(\varphi) := \sup_{f \in \mathcal{F}} \mathbb{E}_S[f(Z)] - \mathbb{E}_T[f(Z)] \tag{17}$$

- catches if the coupling between Z and Y shifts across domains:

$$\mathrm{TSF}(\varphi) := \sup_{\mathbf{f} \in \mathcal{F}_c} \mathbb{E}_S[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_T[Y \cdot \mathbf{f}(Z)] \tag{18}$$

Role of weights

Why designing weights?

- Prediction weighting [Partial Adversarial Domain Adaptation, Cao et al. 2018] ▷ *Estimated labels are used to re-weight the source domain*:

$$w(x) = \frac{p_T(g(z))}{p_S(g(z))}$$

- Entropy conditioning [Conditional Adversarial Domain Adaptation, Long et al. 2018] ▷ *Transfer only confident samples*:

$$w(x) \propto 1 + e^{-H(g(z))} \text{ where } H \text{ is the entropy.}$$

Inductive design of weights

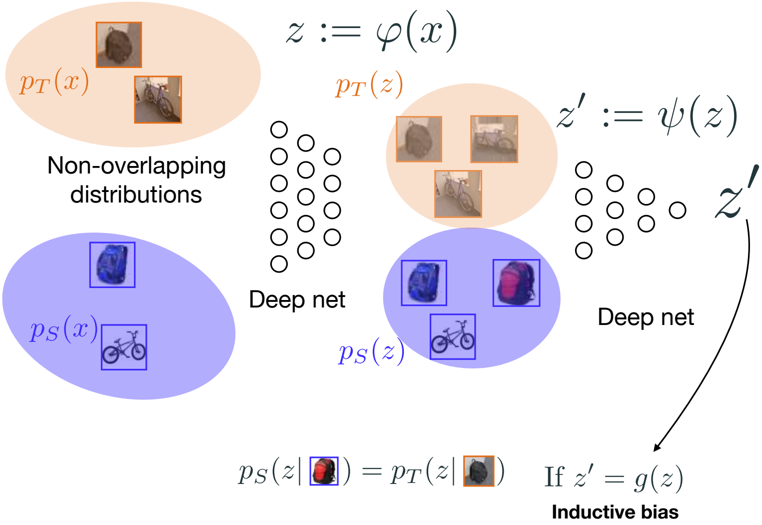
It exists a function $\psi : \mathcal{Z} \to \mathcal{Z}'$, $\psi(z) =: z'$ s.t. w is a function of Z'.

Weights enforces new invariance

If the bound is tight, then,

$$w(z') = \frac{p_T(z')}{p_S(z')} \quad \text{and} \quad p_S(z|z') = p_T(z|z') \tag{19}$$

Role of weights



Detailed view of RUDA

$$\begin{aligned}\varepsilon_T(\tilde{g}\varphi) \leq \frac{\beta}{1-\beta}\{&\varepsilon_{w\cdot S}(g_{w\cdot S}\varphi)\\ &+ 6\cdot \mathrm{INV}(w,\varphi) \triangleright \text{ Set weight s.t. } \mathrm{INV}(w,\varphi) = 0\\ &+ 2\cdot \widehat{\mathrm{TSF}}(w,\varphi,\tilde{g}) \triangleright \text{ Set inductive classifier s.t. } \tilde{g} = g\\ &+ \varepsilon_T(\mathbf{f}_T\varphi)\}\end{aligned}$$

- Set weight s.t. $\mathrm{INV}(w,\varphi) = 0$:

$$w(z) := \frac{p_T(z)}{p_S(z)} = \frac{1-d(z)}{d(z)} \tag{20}$$

  where d is a domain classifier (trained to map 1 in the source domain and 0 in the target domain.)
- Set inductive classifier s.t. $\tilde{g} = g$: $\beta = 1$
  "This is a weak inductive design ($\beta = 1$), thus, theoretical guarantee from bound 4 is not applicable. However, there is empirical evidence that showed that predicted labels help in UDA"

Detailed view of RUDA

$$
\begin{cases}
\theta_\varphi^\star & = \arg\min_{\theta_\varphi} \mathcal{L}_c(\theta_g, \theta_\varphi | \theta_d) + \lambda \cdot \mathcal{L}_{\widehat{\mathrm{TSF}}}(\theta_\varphi, \theta_{\mathbf{d}} | \theta_d, \theta_g) \\
\theta_g & = \arg\min_{\theta_g} \mathcal{L}_c(\theta_g, \theta_\varphi | \theta_d) \\
\theta_d & = \arg\min_{\theta_d} \mathcal{L}_{\mathrm{INV}}(\theta_d | \theta_\varphi)
\end{cases}
\tag{21}
$$

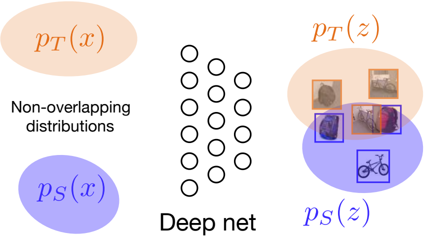
$$
\begin{cases}
\varphi^\star & = \arg\min_{\varphi \in \Phi} \ \varepsilon_{w(\varphi)\cdot S}(g_{w\cdot S}\varphi) + \lambda \cdot \widehat{\mathrm{TSF}}(w, \varphi, g_{w\cdot S}) \\
& \text{such that } w(\varphi) = \arg\min_{w} \mathrm{INV}(w, \varphi)
\end{cases}
\tag{RUDA}
$$

Detailed view of RUDA

$$\mathcal{L}_{\widehat{\mathrm{TSF}}} := \mathbb{E}_S[w(z)Y \cdot \log(\mathsf{d}(z))] + \mathbb{E}_T[g(z) \cdot \log(1 - \mathsf{d}(z))]$$

$$w(z) := \frac{1 - d(z)}{d(z)}$$

$$g(z)$$



$d(z)$ 1 0

$$\mathcal{L}_{\mathrm{INV}} := \mathbb{E}_S[\log(d(z))] + \mathbb{E}_T[\log(1 - d(z))]$$