Adversarially Robust Models may not Transfer Better: Sufficient Conditions for Domain Transferability from the View of Regularization

Xiaojun Xu [*] [1] **Jacky Yibo Zhang** [*] [1] **Evelyn Ma** [1] **Danny Son** [1] **Oluwasanmi Koyejo** [1] **Bo Li** [1]

Abstract

Machine learning (ML) robustness and domain generalization are fundamentally correlated: they essentially concern data distribution shifts under adversarial and natural settings, respectively. On one hand, recent studies show that more robust (adversarially trained) models are more generalizable. On the other hand, there is a lack of theoretical understanding of their fundamental connections. In this paper, we explore the relationship between regularization and domain transferability considering different factors such as norm regularization and data augmentations (DA). We propose a general theoretical framework proving that factors involving the model function class regularization are sufficient conditions for *relative* domain transferability. Our analysis implies that "robustness" is neither necessary nor sufficient for transferability; rather, regularization is a more fundamental perspective for understanding domain transferability. We then discuss popular DA protocols (including adversarial training) and show when they can be viewed as the function class regularization under certain conditions and therefore improve generalization. We conduct extensive experiments to verify our theoretical findings and show several counterexamples where robustness and generalization are negatively correlated on different datasets.

1. Introduction

Domain generalization (or domain transferability) is the task of training machine learning models with data from one or more *source* domains that can be adapted to a *target* domain, often via low-cost fine-tuning. Thus, domain generalization refers to approaches designed to address the *natural data distribution shift* problem (Muandet et al., 2013; Rosenfeld et al., 2021). A wide array of approaches have been proposed to address domain transferability, including fine-tuning the last layer of DNNs (Huang et al., 2018), invariant feature optimization (Muandet et al., 2013), efficient model selection for fine-tuning (You et al., 2019), and optimal transport based domain adaptation (Courty et al., 2016). Understanding domain generalization has emerged as an important task in the machine learning community.

On the other hand, robust machine learning aims to tackle the problem of *adversarial data distribution shift*. Both empirical and certified robust learning approaches have been proposed, such as empirical adversarial training (Madry et al., 2018) and certified defenses based on both deterministic and probabilistic approaches (Cohen et al., 2019; Li et al., 2019; 2021; 2020).

Recent studies (Salman et al., 2020; Utrera et al., 2020) draw a connection between domain transferability and robustness, and suggest that adversarially robust models (i.e., models with good accuracy under adversarial attacks) are more domain transferable. However, a theoretical analysis of their fundamental connections is still lacking, and it is unclear whether robustness is necessary or sufficient. To fill in this gap, this paper aims to answer the following questions: *Is model robustness sufficient or necessary for domain transferability? What are sufficient conditions for domain transferability?*

To answer the first question, our analysis and experiments show that adversarial *robustness is neither sufficient nor necessary* for domain transferability and they can even be negatively correlated. To answer the second question, we first observe that domain transferability is fundamentally a "relative" concept, as it by definition involves two domains, i.e., the source/target domain. With the observation, we propose a general theoretical framework that characterizes sufficient conditions for the *relative* domain transferability from the view of function class regularization. The relative domain transferability, loosely speaking, is the performance of the fine-tuned source model on the target domain relative

[*] Equal contribution [1] University of Illinois at Urbana-Champaign. Correspondence to: Xiaojun Xu <xiaojun3@illinois.edu>, Jacky Yibo Zhang <yiboz@illinois.edu>, Oluwasanmi Koyejo <sanmi@illinois.edu>, Bo Li <lbo@illinois.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

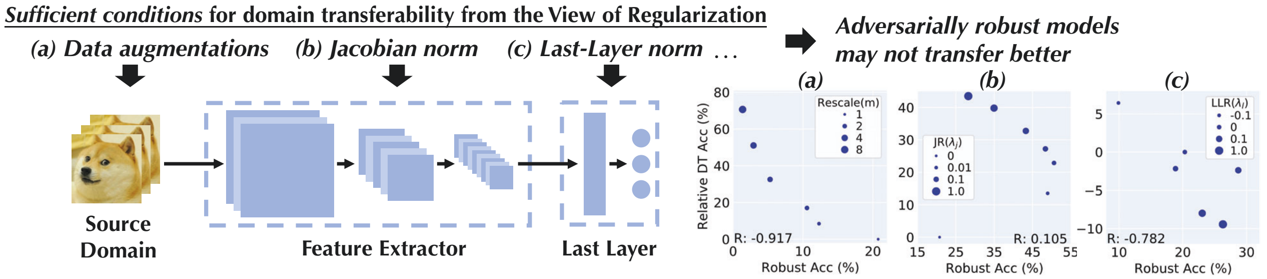

Figure 1. Illustration of robustness and domain transferability in different conditions. We study different augmentation and regularization techniques that can serve as sufficient conditions for domain transferability. We observe that adversarially robust models do not necessarily achieve a better performance in domain transferability and sometimes they are negatively correlated.

to the performance of the source model on the source domain. We then prove an inequality showing that stronger *regularization* on the feature extractor (during the source model training process) implies a better relative domain transferability. We also discuss what data augmentations can be viewed as function class regularization generally. Since adversarial training can be viewed as a regularization under some conditions (Roth et al., 2020; El Ghaoui & Lebret, 1997; Bertsimas & Copenhaver, 2018), our work implies that the regularization effect of adversarial training is a better and more fundamental explanation for the connection between adversarial training and domain transferability.

To verify our theory, we conduct extensive experiments on ImageNet (CIFAR-10 as target domain) and CIFAR-10 (SVHN as target domain) based on different models. We show that regularizations such as norm regularization and certain data augmentations can control the relative and absolute domain transferability, while the robustness and domain transferability can be even negatively correlated with the domain transferability, as illustrated in Fig. 1.

Technical contributions. Our theoretical analysis and empirical findings show that, instead of robustness or adversarial training, regularization is a more fundamental perspective to understand domain transferability. Concretely,

- We show that improving adversarial robustness is neither necessary nor sufficient for improving domain transferability without additional conditions, as shown in Section 2.1.
- We propose a theoretical framework to analyze the sufficient conditions for domain transferability from the view of function class regularization (Section 2.2&2.3). We prove that shrinking the function class of feature extractors during training monotonically decreases a tight upper bound on the relative domain transferability loss. Therefore, it is reasonable to expect that imposing regularization on the feature extractor during training can lead to a better relative domain transferability.
- We provide general analysis on when data augmentations (including adversarial training) can be viewed as regularization. In particular, we verify analysis based on the data augmentations of Gaussian noise, rotation, and translation, as discussed in Section 3.
- We conduct extensive experiments on different datasets and model architectures to verify our theoretical claims (Section 4). We also show counterexamples where adversarial robustness is significantly negatively correlated with domain transferability.

Taken together, our results suggest a more nuanced explanation of the phenomenon that "adversarially trained models transfer better," suggesting instead that adversarial training implies training with regularization, which, in turn, implies better transferability. As a consequence, although adversarial training implies better adversarial robustness, better adversarial robustness does not necessarily imply better transferability.

Related Work. Domain Transferability has been analyzed in different settings. Muandet et al. (2013) present a generalization bound for classification tasks based on the properties of the assumed prior over training environments. Rosenfeld et al. (2021) model domain transferability/generalization as an online game and show that generalizing beyond the convex hull of training environments is NP-hard. Given the complexity of domain transferability analysis, recent empirical studies observe that adversarially trained models transfer better (Salman et al., 2020; Utrera et al., 2020).

Model robustness is an important topic given recent diverse adversarial attacks (Goodfellow et al., 2014; Carlini & Wagner, 2017). These attacks may be launched without access to model parameters (Tu et al., 2019) or even with the model predictions alone (Chen et al., 2020a). Different approaches have been proposed to improve model robustness against adversarial attacks (Yang et al., 2021; Ma et al., 2018; Xiao et al., 2018). Adversarial training has been shown to be effective empirically (Madry et al., 2018; Zhang et al., 2019; Miyato et al., 2018). Some studies have shown that robustness is related to other model characteristics, such as transferability and invertibility (Engstrom et al., 2019; Liang et al., 2020). A recent work (Deng et al., 2021) theoretically analyzes how adversarial training helps transfer learning. Although their proof implicitly depends on regularization, the authors only focus on adversarial training for linear models, while we directly focus on regularization for general models (e.g., DNNs).

2. Sufficient Conditions for Domain Transferability

In this section, we theoretically analyze the problem of domain transferability from the view of regularization and discuss some sufficient conditions for good transferability. All of the proofs are provided in Section A in the appendix.

Notations. We denote the input space as $\mathcal{X}$; the feature space as $\mathcal{Z}$ and the output space as $\mathcal{Y}$. Let the fine-tuning function class be $g \in \mathcal{G}$. Given a feature extractor $f : \mathcal{X} \to \mathcal{Z}$ and a fine-tuning function $g : \mathcal{Z} \to \mathcal{Y}$, the full model is $g \circ f : \mathcal{X} \to \mathcal{Y}$. We denote $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ as the set of distributions on $\mathcal{X} \times \mathcal{Y}$. The loss function on $\mathcal{Y}$ is denoted by $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The population loss function based on data distribution $\mathcal{D} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ and a model $g \circ f$ is defined as

$$\ell_{\mathcal{D}}(g \circ f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(g \circ f(x), y)].$$

In the following, we first provide an example to show that the robustness can be irrelevant to domain transferability and to illustrate why one might investigate domain transferability from the view of regularization.

2.1. A Toy Example: Motivation and Intuition

In this subsection, we construct a simple example where improving adversarial robustness is neither necessary nor sufficient for improving (relative) domain transferability, yet stronger regularization sufficiently improves relative domain transferability. The settings introduced in this subsection are only applied in this subsection.

We consider the case that $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^d$. Given an input $x \in \mathcal{X}$, the ground truth target for the source domain is $y_S(x)$ generated by a function $y_S : \mathbb{R}^m \to \mathbb{R}^d$. Similarly, we define y_T for the target domain. In this example, for simplicity, we neglect the fine-tuning process but directly consider learning a function $f : \mathbb{R}^m \to \mathbb{R}^d$ with a norm $\|\cdot\|$ on $\mathbb{R}^d$. We note that the analysis in this subsection holds with any choice of norm on $\mathbb{R}^d$.

Given the source and target distributions $\mathcal{D}_S, \mathcal{D}_T \in \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$, we consider the case that their marginal distributions on the input space $\mathcal{X}$ are both $\mathcal{D}$, while y_S and y_T could be different. Moreover, we consider the case that the support of the input data distribution $\mathcal{D}$ lies on a low-dimensional manifold $\mathcal{M} \subset \mathcal{X} = \mathbb{R}^m$ such that for $\forall x \in \mathcal{M}$, any Euclidean ball centered at x has non-empty intersection with $\mathbb{R}^m \backslash \mathcal{M}$. Given the distribution $\mathcal{D}$, we define a norm for functions $f : \mathbb{R}^m \to \mathbb{R}^d$ as $\|f\|_{\mathcal{D}} := \mathbb{E}_{x\sim\mathcal{D}}[\|f(x)\|]$, where we view two functions f_1, f_2 as the same if $\|f_1 - f_2\|_{\mathcal{D}} = 0$. Therefore, given a model f, for the source domain and the target domain we consider the respective loss functions as

$$\ell_{\mathcal{D}_S}(f) = \mathbb{E}_{x\sim\mathcal{D}}[\|f(x) - y_S(x)\|] = \|f - y_S\|_{\mathcal{D}},$$
$$\ell_{\mathcal{D}_T}(f) = \mathbb{E}_{x\sim\mathcal{D}}[\|f(x) - y_T(x)\|] = \|f - y_T\|_{\mathcal{D}}.$$

The toy example serves two purposes: (1) supporting the "neither necessary nor sufficient" claim; and (2) motivating the perspective of regularization. For the first purpose, the main intuition is that we can construct a setting where the domain transferability is only evaluated on a low-dimensional manifold while the adversarial robustness is only evaluated off the manifold. In such cases, a model having better adversarial robustness does not imply it has better domain transferability, and similarly a model having better domain transferability does not imply it has better adversarial robustness. For the second purpose, as illustrated in Figure 2, regularization is related to the domain transferability in this toy example. This motivates the general study of the relationship between regularization and domain transferability in Section 2.2.

Robustness is neither necessary nor sufficient for domain transferability. We may see the relation between adversarial robustness and domain transferability in this example as follows. Given a source model $f^{\mathcal{D}_S} : \mathbb{R}^m \to \mathbb{R}^d$, we consider the adversarial loss on an input $x \in \mathcal{M}$, i.e.,

$$\ell_{adv}(x; f^{\mathcal{D}_S}) := \max_{\delta:\|\delta\|_2\leq\epsilon} \ell(f^{\mathcal{D}_S}(x+\delta), y_S(x)), \tag{1}$$

as an indicator of its robustness on the input x on the source domain. The lower the adversarial loss, the better the robustness. We can see that both the regular loss functions $\ell_{\mathcal{D}_S}(f^{\mathcal{D}_S})$ and $\ell_{\mathcal{D}_T}(f^{\mathcal{D}_S})$ only evaluate $f^{\mathcal{D}_S}$ on the low-dimensional manifold $\mathcal{M}$. Therefore, an adversarial perturbation $\delta \in \mathbb{R}^m$ could make $x + \delta \notin \mathcal{M}$ if the loss value is sufficiently high in $\{x + \delta \mid \|\delta\|_2 \leq \epsilon\}\backslash\mathcal{M}$. As a result, in such cases the adversarial loss $\ell_{adv}(x; f^{\mathcal{D}_S})$ could be arbitrarily high without affecting either the source domain performance $\ell_{\mathcal{D}_S}(f^{\mathcal{D}_S})$ or the target domain performance $\ell_{\mathcal{D}_T}(f^{\mathcal{D}_S})$, i.e., without affecting their transferability. This implies that improving adversarial robustness is neither necessary nor sufficient for improving domain transferability.

The toy example illustrates that robustness can be irrelevant to domain transferability, and then the question one may naturally ask is "what may have a stronger relevance to domain transferability?" To provide the intuition that regularization may be the key, we make the following analysis using the same toy example.

Intuition on why regularization matters. Denoting a function space $\mathcal{F} = \{f : \mathbb{R}^m \to \mathbb{R}^d \mid \|f\|_{\mathcal{D}} < \infty\}$, we

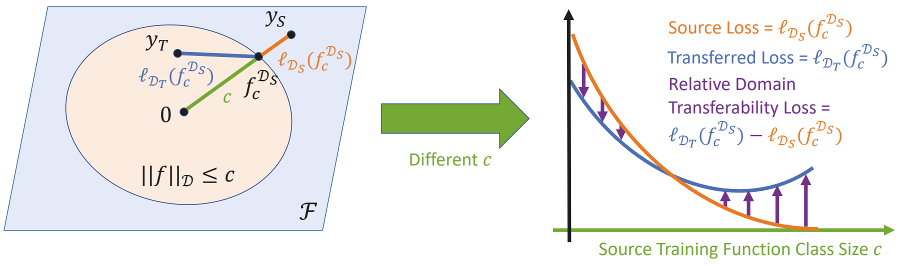
y_S
y_T
$\ell_{\mathcal{D}_T}(f_c^{\mathcal{D}_S})$
$\ell_{\mathcal{D}_S}(f_c^{\mathcal{D}_S})$
$f_c^{\mathcal{D}_S}$
c
0
$||f||_{\mathcal{D}} \leq c$
$\mathcal{F}$
Different c
Source Loss = $\ell_{\mathcal{D}_S}(f_c^{\mathcal{D}_S})$
Transferred Loss = $\ell_{\mathcal{D}_T}(f_c^{\mathcal{D}_S})$
Relative Domain Transferability Loss = $\ell_{\mathcal{D}_T}(f_c^{\mathcal{D}_S}) - \ell_{\mathcal{D}_S}(f_c^{\mathcal{D}_S})$
Source Training Function Class Size c

Figure 2. The left figure illustrates the example in the function space $\mathcal{F}$ given a regularization parameter c. The right figure shows the relations between domain transferability and the c. In this example, the stronger the regularization effect (smaller c) is, the lower the relative domain transferability loss is (violet arrow), and the better the relative domain transferability is.

assume $y_S, y_T \in \mathcal{F}$ such that we can compare f, y_S, y_T in the same space. Therefore, given $c > 0$ as a regularization parameter, we define the domain transferability problem as:

Learning a source model:

$$f_c^{\mathcal{D}_S} \in \arg\min_{f \in \mathcal{F}} \ell_{\mathcal{D}_S}(f), \quad \text{s.t. } \|f\|_{\mathcal{D}} \leq c; \tag{2}$$

Testing on a target domain: $\quad \ell_{\mathcal{D}_T}(f_c^{\mathcal{D}_S}),$

where the minimizer is $f_c^{\mathcal{D}_S} := y_S \min\{1, \frac{c}{\|y_S\|_{\mathcal{D}}}\}$, the source domain loss is $\ell_{\mathcal{D}_S}(f) = \|f - y_S\|_{\mathcal{D}}$, and the target domain loss is $\ell_{\mathcal{D}_T}(f) = \|f - y_T\|_{\mathcal{D}}$. We prove in Proposition 2.1 that $f_c^{\mathcal{D}_S}$ is indeed a minimizer of equation 2.

Considering the relation between (relative) domain transferability and the regularization parameter c, we have an interesting finding. An illustration of the finding is shown in Figure 2, and a more formal statement is provided in Proposition 2.1. As we can see, the relation between regularization and domain transferability is clear if we consider the domain transferability in a "relative" way, i.e., the loss value on the target domain minus the loss value on the source domain. A formal definition of the relative transferability loss is deferred to Definition 2.2 in the next subsection.

Proposition 2.1. *Given the toy example problem defined in Section 2.1, $f_c^{\mathcal{D}_S}$ is a minimizer of equation 2. If $c \geq c' \geq 0$, then the relative domain transferability loss $\ell_{\mathcal{D}_T}(f_c^{\mathcal{D}_S}) - \ell_{\mathcal{D}_S}(f_c^{\mathcal{D}_S}) \geq \ell_{\mathcal{D}_T}(f_{c'}^{\mathcal{D}_S}) - \ell_{\mathcal{D}_S}(f_{c'}^{\mathcal{D}_S})$.*

As we can see from this toy example, robustness is neither necessary nor sufficient to characterize domain transferability. However, there is a monotone relation between the regularization strength and the relative domain transferability loss. Although the above proposition is derived specifically for the toy example, similar behavior is also observed in our experiments. Naturally, these findings motivate the study of the connections between the regularization of the training process and domain transferability in general, as we consider next.

2.2. Upper Bound of Relative Domain Transferability

In this subsection, we consider the general transferability problem with fine-tuning. We prove that there is a monotone decreasing relationship between the regularization strength and a tight upper bound on the relative domain transferability loss. Given a training algorithm A, it takes a data distribution $\mathcal{D}$ and outputs a feature extractor $f_A^{\mathcal{D}} \in \mathcal{F}_A$ chosen from a function class $\mathcal{F}_A$ as well as a fine-tuning function $g_A^{\mathcal{D}} \in \mathcal{G}$. First, we formally define the relative domain transferability loss.

Definition 2.2 (Relative Domain Transferability Loss)**.** Given the training algorithm A and a pair of distributions $\mathcal{D}_S, \mathcal{D}_T \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, the relative domain transferability loss between $\mathcal{D}_S, \mathcal{D}_T$ is defined to be the difference of fine-tuned losses, i.e.,

$$\tau(A; \mathcal{D}_S, \mathcal{D}_T) := \inf_{g \in \mathcal{G}} \ell_{\mathcal{D}_T}(g \circ f_A^{\mathcal{D}_S}) - \ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S}).$$

As we can see, when $\ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S})$ is the same, smaller $\tau(A; \mathcal{D}_S, \mathcal{D}_T)$ means the better performance on the target domain.

Another perspective of Definition 2.2 is that $\inf_{g\in\mathcal{G}} \ell_{\mathcal{D}_T}(g \circ f_A^{\mathcal{D}_S}) = \ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S}) + \tau(A; \mathcal{D}_S, \mathcal{D}_T)$. From this perspective, the transferred loss is the source loss plus an additional term to be upper bounded by a certain distance metric between the source and target distributions – as is common in the literature of domain adaptation (e.g., (Ben-David et al., 2007; Zhao et al., 2019)). The key question of the "distance metric" remains unanswered. To this end, we propose the following.

Definition 2.3 ($(\mathcal{G}, \mathcal{F})$-pseudometric)**.** Given a fine-tuning function class $\mathcal{G}$, a feature extractor function class $\mathcal{F}$ and distributions $\mathcal{D}_S, \mathcal{D}_T \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, the $(\mathcal{G}, \mathcal{F})$-pseudometric between $\mathcal{D}_S, \mathcal{D}_T$ is

$$d_{\mathcal{G},\mathcal{F}}(\mathcal{D}_S, \mathcal{D}_T) := \sup_{f \in \mathcal{F}} |\inf_{g \in \mathcal{G}} \ell_{\mathcal{D}_S}(g \circ f) - \inf_{g \in \mathcal{G}} \ell_{\mathcal{D}_T}(g \circ f)|.$$

Since the fine-tuning function class is usually simple and fixed, we will use $d_{\mathcal{F}}$ as an abbreviation when $\mathcal{G}$ is clear.

It can be easily verified that $d_{\mathcal{G},\mathcal{F}}$ is a pseudometric that measures the *distance* between two distributions, as shown in the following proposition.

Proposition 2.4. $d_{\mathcal{G},\mathcal{F}}(\cdot,\cdot) : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \times \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}_+$ *satisfies the following properties.*

1. *(Symmetry)* $d_{\mathcal{G},\mathcal{F}}(\mathcal{D}_S, \mathcal{D}_T) = d_{\mathcal{G},\mathcal{F}}(\mathcal{D}_T, \mathcal{D}_S)$.
2. *(Triangle Inequality) For* $\forall \mathcal{D}' \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, *we have* $d_{\mathcal{G},\mathcal{F}}(\mathcal{D}_S, \mathcal{D}_T) \le d_{\mathcal{G},\mathcal{F}}(\mathcal{D}_S, \mathcal{D}') + d_{\mathcal{G},\mathcal{F}}(\mathcal{D}', \mathcal{D}_T)$.
3. *(Weak Zero Property) For* $\forall \mathcal{D} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$: $d_{\mathcal{G},\mathcal{F}}(\mathcal{D}, \mathcal{D}) = 0$.

The motivation of the $(\mathcal{G},\mathcal{F})$-pseudometric comes from the following observations. We want to study what factors affect how a source model transfers to the target domain. The obvious factor is the difference between the two domains. But the function class where the model is trained from is also an important factor (e.g., the example in Section 2.1). Note that the proposed $(\mathcal{G},\mathcal{F})$-pseudometric is both a complexity measure of the model function class and a distance measure of two distributions. Given a certain fixed function class, the $(\mathcal{G},\mathcal{F})$-pseudometric can serve as a distance measure related to the Wasserstein distance or the total variance distance. In proposition A.3 in the appendix, we show that, if the loss function class is Lipschitz, then the $(\mathcal{G},\mathcal{F})$-pseudometric between $\mathcal{D}_S$ and $\mathcal{D}_T$ is upper bounded by the product of the Lipschitz constant and the Wasserstein distance between $\mathcal{D}_S$ and $\mathcal{D}_T$. Moreover, in proposition A.4 in the appendix, we show that the total variation distance upper bounds the $(\mathcal{G},\mathcal{F})$-pseudometric if we are working in the realm of multi-class classification and the loss function is the 0-1 loss.

The major difference of the $(\mathcal{G},\mathcal{F})$-pseudometric with existing metrics for domain transfer (Ben-David et al., 2010; Mansour et al., 2009; Acuna et al., 2021; Zhao et al., 2019) is that the proposed $(\mathcal{G},\mathcal{F})$-pseudometric is more general. Concretely, the aforementioned work only considers the distributions on the input space $\mathcal{X}$, while we consider both the input space and the output space, i.e, $\mathcal{X} \times \mathcal{Y}$. This difference enables us to consider the fine-tuning process, which is important and widely applied in practice.

In this section, we consider a fixed fine-tuning function class $\mathcal{G}$ and feature extractor function class $\mathcal{F}_A$ given by the training algorithm A. Thus, we denote $d_{\mathcal{G},\mathcal{F}}$ as $d_{\mathcal{F}_A}$ for the remainder of the paper. With the definition of $d_{\mathcal{F}_A}$, we can derive the following result which provides justification for the regularization perspective.

Theorem 2.5. *Given a training algorithm* A, *for* $\forall \mathcal{D}_S, \mathcal{D}_T \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ *we have*

$$\tau(A; \mathcal{D}_S, \mathcal{D}_T) \le d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T), \quad \textit{or equivalently,}$$
$$\inf_{g\in\mathcal{G}} \ell_{\mathcal{D}_T}(g \circ f_A^{\mathcal{D}_S}) \le \ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S}) + d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T).$$

Interpretation: As we can see, the above theorem provides sufficient conditions for good domain transferability. There is a monotone relation between the regularization strength and $d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T)$, i.e., the upper bound on the relative domain transferability loss $\tau(A; \mathcal{D}_S, \mathcal{D}_T)$. More explicitly, if a training algorithm A' has $\mathcal{F}_{A'} \subseteq \mathcal{F}_A$, then $d_{\mathcal{F}_{A'}}(\mathcal{D}_S, \mathcal{D}_T) \le d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T)$. Moreover, small $d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T)$ implies good relative domain transferability. From this perspective, we can see that we need both small $d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T)$ and small source loss $\ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S})$ to guarantee good absolute domain transferability. Note that there is a possible trade-off, i.e., with $\mathcal{F}_A$ being smaller, $d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T)$ decreases but possibly $\ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S})$ increases due to the limited power of $\mathcal{F}_A$. On the other hand, there may not be such trade-off if $\mathcal{D}_S$ and $\mathcal{D}_T$ are close enough such that $d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T)$ is small.

To make the upper bound more meaningful, we need to study its tightness.

Theorem 2.6. *Given any source distribution* $\mathcal{D}_S \in \mathcal{P}_{\mathcal{X}\times\mathbb{R}^d}$, *any fine-tuning function class* $\mathcal{G}$ *where* $\mathcal{G}$ *includes the zero function, we assume the training algorithm* A *is optimal, i.e.,* $\ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S}) = \inf_{g\in\mathcal{G}, f\in\mathcal{F}_A} \ell_{\mathcal{D}_S}(g \circ f)$. *We assume some properties of the loss function* $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$: *it is differentiable and strictly convex w.r.t. its first argument;* $\ell(y, y) = 0$ *for any* $y \in \mathbb{R}^d$; *and* $\lim_{r\to\infty} \inf_{y:\|y\|_2=r} \ell(\vec{0}, y) = \infty$, *where* $\vec{0}$ *is the zero vector. Then, given any distribution* $\mathcal{D}^{\mathcal{X}}$ *on* $\mathcal{X}$, *there exist some distributions* $\mathcal{D}_T \in \mathcal{P}_{\mathcal{X}\times\mathbb{R}^d}$ *with its marginal on* $\mathcal{X}$ *being* $\mathcal{D}^{\mathcal{X}}$ *such that*

$$\tau(A; \mathcal{D}_S, \mathcal{D}_T) = d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T), \quad \textit{or equivalently,}$$
$$\inf_{g\in\mathcal{G}} \ell_{\mathcal{D}_T}(g \circ f_A^{\mathcal{D}_S}) = \ell_{\mathcal{D}_S}(g_A^{\mathcal{D}_S} \circ f_A^{\mathcal{D}_S}) + d_{\mathcal{F}_A}(\mathcal{D}_S, \mathcal{D}_T).$$

Interpretation: In the above theorem, we show that given any $A, \mathcal{D}_S$, and the marginal $\mathcal{D}^{\mathcal{X}}$, there exist some conditional distributions of $y|x$ such that by composing it with the given $\mathcal{D}^{\mathcal{X}}$ we have a distribution $\mathcal{D}_T$ where the equality holds in Theorem 2.5. The optimality assumption on the training algorithm is mild, as it is common for modern neural networks to achieve considerably low loss. Nonetheless, a generalized version of the theorem is provided as Theorem A.7 in the appendix which works with *any* training algorithm. Alternative form of the tightness analysis is discussed immediately after the proof of Theorem A.7.

Therefore, we prove that stronger regularization on the feature extractor implies a decreased tight upper bound on the relative transferability loss. For a cleaner presentation, the analysis so far does not consider the potential influence from finite samples which for sure affects domain generalization. In the next subsection, we investigate the proposed theory on relative transferability with finite samples.

2.3. Generalization Upper Bound of the Relative Domain Transferability

For a distribution $\mathcal{D} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, we denote its empirical distribution with n samples as $\widehat{\mathcal{D}}^n$. That being said,

$$\begin{aligned} \ell_{\widehat{\mathcal{D}}^n}(g \circ f) &= \mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}^n}[\ell(g \circ f(x), y)] \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(g \circ f(x_i), y_i), \end{aligned}$$

where (x_i, y_i) are i.i.d. samples from $\mathcal{D}$. Therefore, given two distributions $\mathcal{D}_S, \mathcal{D}_T \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, the empirical $(\mathcal{G}, \mathcal{F})$-pseudometric between them is $d_{\mathcal{G},\mathcal{F}}(\widehat{\mathcal{D}}_S^n, \widehat{\mathcal{D}}_T^n)$.

Note that $d_{\mathcal{G},\mathcal{F}}$ is not only a pseudometric of distributions, but also a complexity measure, and we will first connect it with the Rademacher complexity.

Definition 2.7 (Empirical Rademacher Complexity (Bartlett & Mendelson, 2002; Koltchinskii, 2001))**.** Denote the loss function class induced by $\mathcal{G}, \mathcal{F}$ as

$$\mathcal{L}_{\mathcal{G},\mathcal{F}} := \{h_{g,f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \mid g \in \mathcal{G}, f \in \mathcal{F}\},$$

where $h_{g,f}(x, y) := \ell(g \circ f(x), y)$.

Given an empirical distribution $\widehat{\mathcal{D}}^n$ (i.e., n data samples), the Rademacher complexity of it is

$$\mathrm{Rad}_{\widehat{\mathcal{D}}^n}(\mathcal{L}_{\mathcal{G},\mathcal{F}}) := \frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{h\in\mathcal{L}_{\mathcal{G},\mathcal{F}}} \sum_{i=1}^n \xi_i h(x_i, y_i)\right],$$

where $\boldsymbol{\xi} \in \mathbb{R}^n$ are Rademacher variables, i.e., each ξ_i is i.i.d. uniformly distributed on $\{-1, 1\}$.

We can see that if there is a $\mathcal{F}' \subseteq \mathcal{F}$, then $\mathrm{Rad}_{\widehat{\mathcal{D}}^n}(\mathcal{L}_{\mathcal{G},\mathcal{F}'}) \leq \mathrm{Rad}_{\widehat{\mathcal{D}}^n}(\mathcal{L}_{\mathcal{G},\mathcal{F}})$. With the above definitions, we have the following lemma connecting the $(\mathcal{G}, \mathcal{F})$-pseudometric to Rademacher complexity.

Lemma 2.8. *Assuming the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, c]$, given any distribution $\mathcal{D} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ and $\forall\delta > 0$, with probability $\geq 1 - \delta$ we have*

$$d_{\mathcal{G},\mathcal{F}}(\mathcal{D}, \widehat{\mathcal{D}}^n) \leq 2\mathrm{Rad}_{\widehat{\mathcal{D}}^n}(\mathcal{L}_{\mathcal{G},\mathcal{F}}) + 3c\sqrt{\frac{\ln(4/\delta)}{2n}}.$$

Therefore, denoting again $d_{\mathcal{F}_A}$ as $d_{\mathcal{G},\mathcal{F}_A}$, the empirical version of Theorem 2.5 is as follows.

Theorem 2.9. *Assuming the loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to [0, c]$, given $\forall\mathcal{D}_S, \mathcal{D}_T \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, for $\forall\delta > 0$ with probability $\geq 1 - \delta$ we have*

$$\begin{aligned} \tau(A; \widehat{\mathcal{D}}_S^n, \mathcal{D}_T) \leq{}& d_{\mathcal{F}_A}(\widehat{\mathcal{D}}_S^n, \widehat{\mathcal{D}}_T^n) + 2\mathrm{Rad}_{\widehat{\mathcal{D}}_T^n}(\mathcal{L}_{\mathcal{G},\mathcal{F}_A}) \\ &+ 4\mathrm{Rad}_{\widehat{\mathcal{D}}_S^n}(\mathcal{L}_{\mathcal{G},\mathcal{F}_A}) + 9c\sqrt{\frac{\ln(8/\delta)}{2n}}. \end{aligned}$$

Interpretation: We can see that a smaller feature extractor function class $\mathcal{F}_A$ implies both a smaller $d_{\mathcal{F}_A}$ and the Rademacher complexity. Therefore, the monotone relation between the regularization strength and the upper bound on the relative domain transferability loss also holds for the empirical settings.

The proposed theoretical analysis suggests that regularization may be a fundamental perspective to understand domain transferability. Other than explicit regularization, empirically we find that the transferability is also related to the use of certain data augmentation and adversarial training. Can we explain such phenomena from the view of regularization again? We discuss this question in the next section.

3. When Can Data Augmentation be Viewed as Regularization?

In this section, we discuss the connections between data augmentation (DA) and regularization. We present the results and their interpretation in this section, while deferring the detailed discussion and comparisons with related work to Section B in the appendix.

General settings. We consider the fine-tuning function $g : \mathbb{R}^d \to \mathbb{R}$ as a linear layer, which will be concatenated to the feature extractor $f : \mathbb{R}^m \to \mathbb{R}^d$. Given a model $g \circ f$, we use the squared loss $\ell(g \circ f(x), y) = (g \circ f(x) - y)^2$, and accordingly apply second-order Taylor expansion to the objective function to study the effect of data augmentation.

DA categories. We discuss two categories of DA, the *feature-level DA* and the *data-level DA*. The feature-level DA (Wong et al., 2016; DeVries & Taylor, 2017) requires the transformation to be performed in the learned feature space: given a data sample $x \in \mathbb{R}^m$ and a feature extractor f, the augmented feature is $W_\star f(x) + b_\star$ where $W_\star \in \mathbb{R}^{d\times d}, b_\star \in \mathbb{R}^d$ are sampled from a distribution. On the other hand, the data-level DA requires the transformation to be performed in the input space: given a data sample x, the augmented sample is $W_\star x + b_\star$ where $W_\star \in \mathbb{R}^{m\times m}, b_\star \in \mathbb{R}^m$ are sampled from a distribution.

Intuition on sufficient conditions. For either the feature-level or the data-level DA, the intuitions given by our analysis are similar. Our results (Theorem B.1&B.2) suggest that the following conditions indicate regularization effects of a data augmentation: 1) $\mathbb{E}_{W_\star}[W_\star] = \mathbb{I}$; 2) $\mathbb{E}_{b_\star}[b_\star] = \vec{0}$; 3) $W_\star$ and $b_\star$ are independent, where $\mathbb{I}$ is the identity matrix and $\vec{0}$ is the zero vector; 4) $W_\star$ is not a constant if it is the feature-level DA; 5) DA is of a small magnitude if it is the data-level DA.

Empirical verification. Combining with Theorem 2.9, it suggests that DA satisfying the conditions above may improve the relative domain transferability. In fact, it matches the empirical observations in Section 4. Concretely, ***1) Gaussian noise*** *satisfies* the four conditions, and empirically the Gaussian noise improves domain transferability while robustness decreases a bit (Figure 5); ***2) Rotation***, which rotates input image with a predefined fixed angle with predefined fixed probability, *violates* $\mathbb{E}_{W_\star}[W_\star] = \mathbb{I}$, and empirically the rotation barely affects domain transferability (Figure 7); ***3) Translation***, which moves the input image for a predefined distance along a pre-selected axis with fixed probability, *violates* $\mathbb{E}_{b_\star}[b_\star] = \vec{0}$, and empirically the translation distance barely co-relates to the domain transferability (Figure 7).

Adversarial training. It is known that adversarial training, a special kind of data augmentation, can be viewed as regularization in some scenarios (Roth et al., 2020). We further prove that, under certain conditions, adversarial training reduces the size of the feature extractors function class during training (see Section C for details). Therefore, our theoretical analysis implies that adversarial training helps domain transferability from its regularization effect.

4. Evaluation

4.1. Experimental Setting

Source model training. We train our model on two source domains: CIFAR-10 and ImageNet. Unless specified, we will use the training settings as follows[1]. For CIFAR-10, we train the model with 200 epochs using the momentum SGD optimizer with momentum 0.9, weight decay 0.0005, an initial learning rate 0.1 which decays by a factor of 10 at the 100-th and 150-th epoch. For ImageNet, we train the model with 90 epochs using the momentum SGD optimizer with momentum 0.9, weight decay 0.0001, an initial learning rate 0.1 which decays by a factor of 10 at the 30-th and 60-th epoch. We use the standard cross-entropy loss denote as $L_{CE}(h_s, x, y)$, where $h_s = g_s \circ f$ is the trained model and x, y are the input and label respectively. For both tasks, we use ResNet-18 as the model architecture. We provide results of other model structures in Appendix D.3.

Model robustness evaluation. To evaluate the model robustness on the source domain, we will show the model accuracy under adversarial attack. We follow the evaluation setting in (Ilyas et al., 2019) and perform the PGD attack with 20 steps using $\epsilon = 0.25$. This empirical robust accuracy reflects how well the model performs under adversarial attack, which is the adversarial loss as in equation 1 if we view $\ell(\cdot, \cdot)$ as the 0-1 loss between prediction and ground truth. We also provide robustness evaluation with AutoAttack in Appendix D.4.

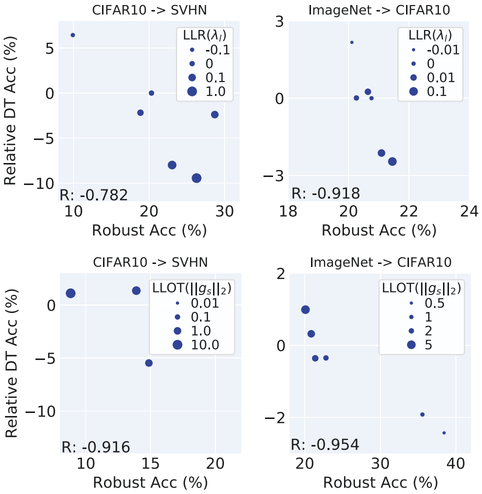

Figure 3. Relationship between robustness and transferability under different norms of last layer, via training with last-layer regularization (LLR) and last-layer orthogonalization (LLOT)

Domain transferability. We evaluate the transferability from CIFAR-10 to SVHN and from ImageNet to CIFAR-10. For the ImageNet, we focus on CIFAR as the target domain, since it is the domain that is the most positively correlated with robustness as shown in (Salman et al., 2020). We evaluate the fixed-feature transfer where only the last fully-connected layer is fine-tuned following our theoretical framework. We fine-tune the last layer with 40 epochs using SGD with momentum 0.9, weight decay 0.0005, an initial learning rate 0.01 which decays by a factor of 10 at the 20-th and 30-th epoch. To mitigate the impact of benign accuracy, we evaluate the *relative domain transfer accuracy* (DT Acc) as follows. Let acc_{src} and acc_{tgt} be the accuracy of the fine-tuned model on the source and target domain, and acc^v_{src} and acc^v_{tgt} be the accuracy of vanilla model (*i.e.*, models trained with standard settings) on source and target domain, then the relative DT accuracy is defined as:

$$\text{DT Acc} = (acc_{tgt} - acc_{src}) - (acc^v_{tgt} - acc^v_{src}).$$

Note that by definition, we can directly use $acc_{tgt} - acc_{src}$ as the relative accuracy. We use a relative score ($acc^v_{tgt} - acc^v_{src}$) so that the positive/negative values reflect the comparison with the vanilla-trained model. We also provide the results of absolute DT accuracy in Appendix D.1.

4.2. Relationship between Robustness and Transferability Under Controllable Conditions

We train the model under different controllable conditions to validate our analysis. In particular, we train the methods by controlling different regularization or data augmentations to evaluate the model robustness and transferability. We

[1] These settings are inherited from the standard training algorithms for CIFAR-10 (`https://github.com/kuangliu/pytorch-cifar`) and ImageNet (`https://github.com/pytorch/examples/tree/master/imagenet`).

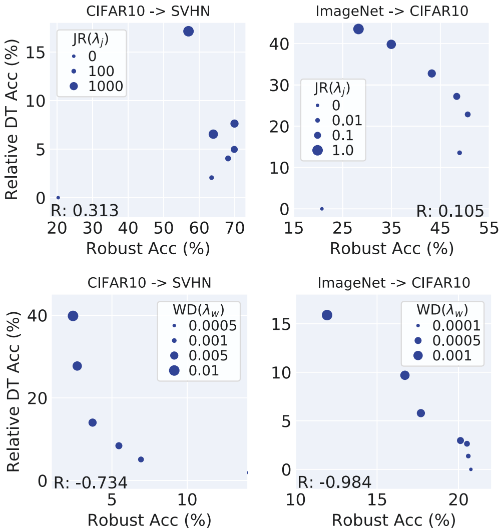

Figure 4. Relationship between robustness and transferability when we regularize the feature extractor with Jacobian Regularization (JR) and weight decay (WD).

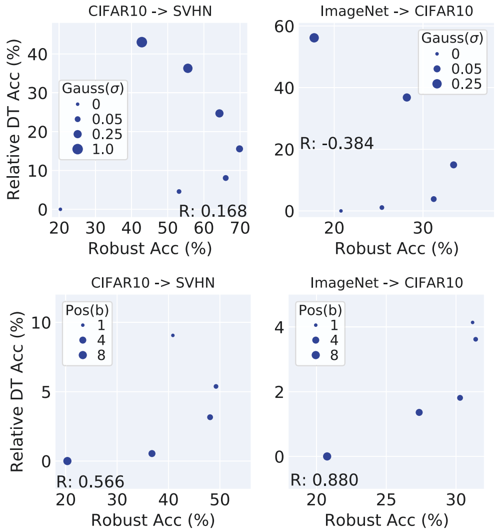

Figure 5. Relationship between robustness and transferability when we use Gaussian noise (*Gauss*) and posterize (*Pos*) as data augmentations.

emphasize that our goal is to identify conditions for domain transferability, rather than proposing methods to achieve the state-of-the-art transferable models. Nevertheless, we do show in Appendix D.2 that with basic regularization the model can achieve better absolute transferability than vanilla trained or adversarially trained models in some cases.

Controlling the last-layer norm. As shown in our theory, (relative) domain transferability is related to the regularization of feature extractors. Here we regularize the transferability by controlling the last-layer norm g_s. Intuitively, when we force the norm of g_s to be big during training, the corresponding norm of f will be regularized to be small. We use two approaches to control the last-layer norm:

- Last-layer regularization (LLR): we impose a strong l2-regularizer with parameter λ_l specifically on the weight of g_s and therefore our training loss becomes: $L_{LLR}(h_s, x, y) = L_{CE}(h_s, x, y) + \lambda_l \cdot ||g_s||_F$, where $||g_s||_F$ is the frobenius norm of the weight matrix of g_s.
- Last-layer orthogonal training (LLOT): we directly control the l2-norm of g_s with orthogonal training ((Huang et al., 2020)). The orthogonal training will enforce the weight to become a 1-norm matrix and we multiply a constant to obtain the desired norm $||g_s||_2$.

The result of LLR and LLOT are shown in Figure 3. We observe that when we regularize the norm of the last layer to be large (i.e. smaller λ in LLR and larger $||g_s||_2$ in LLOT), the relative domain transferability will increase while the model robustness will decrease (their negative correlation is significant with Pearson's coefficient around -0.9). This is because the larger last layer norm will produce a feature extractor f with a smaller norm, which, according to our analysis, leads to a better relative domain transferability. On the other hand, the model $g_s \circ f$ will have a larger norm and therefore becomes less robust under adversarial attacks.

Controlling the norm of feature extractor. We directly regularize the feature extractor f and check the impact on the (relative) domain transferability. We implement two regularization as follows:

- Jacobian regularization (JR): we follow the approach in (Hoffman et al., 2019) to apply JR on the feature extractor. Given model $h_s = g_s \circ f$, the training loss becomes: $L_{JR}(g_s \circ f, x, y) = L_{CE}(g_s \circ f, x, y) + \lambda_j \cdot ||J(f, x)||_F^2$, where $J(f, x)$ denotes the Jacobian matrix of f on x and $|| \cdot ||_F$ is the frobenius norm.
- Weight Decay (WD): we impose weight decay with factor λ_w on the feature extractor f during training. This is equivalent to imposing l2-regularizer with factor λ_w on the feature extractor (excluding the last layer).

The results under JR and WD are shown in Figure 4. We observe that with larger regularization on the feature extractor, the model shows higher relative domain transferability, which matches our analysis. Meanwhile, the robustness decreases significantly with a large regularizer. This is because a large regularization will harm the model performance on the source domain and lead to low model robustness.

Noise-dependent data augmentation. As shown in Section 3, certain data augmentation can be viewed as a type of regularization during training and thus affects the (relative) domain transferability. Here we consider both noise

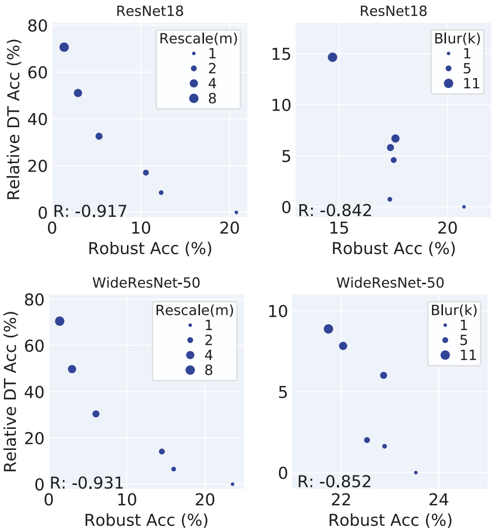

Figure 6. Relationship between robustness and transferability on ImageNet when we use rescale and blur as data augmentations.

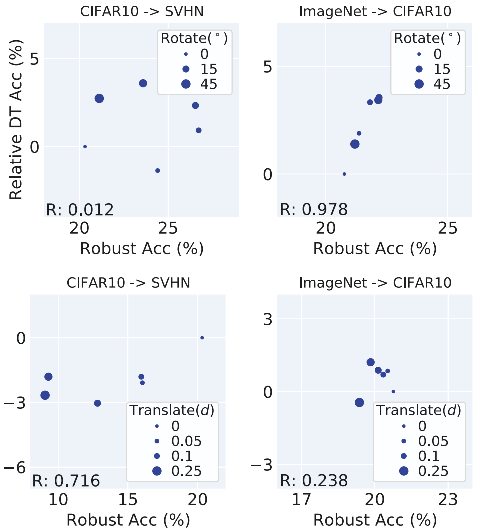

Figure 7. Relationship between robustness and transferability when we use rotation and translation as data augmentations, which violate the sufficient condition for regularization. We cannot see any obvious trend for such augmentations.

dependent and independent data augmentations. For the noise-dependent case, We include two augmentations:

- Gaussian Noise data augmentation (*Gauss*): we add zero-mean Gaussian noise with variance σ^2 to the input image.
- Posterize (*Pos*): we truncate each channel of one pixel value into b bits (originally they are 8 bits).

The results of *Gauss* and *Pos* are shown in Figure 5. We observe that the relative domain transferability of the trained models improves with greater data augmentation, matching our theory. The robustness also benefits from a small data augmentation but decreases when it becomes large.

Resolution-related (noise-independent) data augmentation. Specifically, for ImageNet to CIFAR-10 transferability, we consider two resolution-related data augmentations. The intuition is that when the target domain has a lower resolution than the source domain (ImageNet is 224×224 while CIFAR-10 is 32×32), the data augmentations that down-sample the inputs during the training on the source domain will help transferability. We consider the below resolution-related augmentations:

- Rescale: we rescale the input to be m times smaller (*i.e.*, shape ImageNet as $(224/m) \times (224/m)$) and then rescale them back to the original size.
- Blur: we apply Gaussian blurring with kernel size k on the input. The Gaussian kernel is created with a standard deviation randomly sampled from $[0.1, 2.0]$.

The corresponding results are shown in Figure 6. The experiments are evaluated only for ImageNet to CIFAR-10, and we include the results of both ResNet18 (the default model) and WideResNet50. We can see that the data augmentations help with relative domain transferability to the target domain, although the robustness on the source domain decreases since these augmentations do not relate to robustness operations

4.3. Other Data Augmentations

In addition, we study rotation and translation, the two data augmentations that violate the sufficient condition for regularization as we discussed in Section 3. The result is shown in Figure 7. We observe that these augmentations do not have an obvious impact on domain transferability, which is consistent with our theoretical analysis.

5. Conclusions

In this work, we theoretically analyze the sufficient conditions for (relative) domain transferability based on the view of function class regularization. We also conduct experiments to verify our claims and observe some counterexamples that show negative correlations between robustness and domain transferability. These results would contribute to a better understanding of the domain generalization.

Acknowledgement

This work is partially supported by NSF 1910100, NSF 2046795, NSF 1909577, NSF 1934986, NSF CNS 2046726, NIFA award 2020-67021-32799, C3 AI, and the Alfred P. Sloan Foundation.

References

Acuna, D., Zhang, G., Law, M. T., and Fidler, S. f-domain-adversarial learning: Theory and algorithms. In *ICML*, 2021.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bertsimas, D. and Copenhaver, M. S. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931–942, 2018.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Carratino, L., Cissé, M., Jenatton, R., and Vert, J.-P. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

Chen, J., Jordan, M. I., and Wainwright, M. J. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pp. 1277–1294. IEEE, 2020a.

Chen, S., Dobriban, E., and Lee, J. H. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020b.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pp. 1528–1537. PMLR, 2019.

Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., and Zou, J. Adversarial training helps transfer learning via better representations. *arXiv preprint arXiv:2106.10189*, 2021.

DeVries, T. and Taylor, G. W. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

El Ghaoui, L. and Lebret, H. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064, 1997.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., and Madry, A. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Greenewald, K., Gu, A., Yurochkin, M., Solomon, J., and Chien, E. k-mixup regularization for deep learning via optimal transport. *arXiv preprint arXiv:2106.02933*, 2021.

Hernández-García, A. and König, P. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018a.

Hernández-García, A. and König, P. Further advantages of data augmentation on convolutional neural networks. In *International Conference on Artificial Neural Networks*, pp. 95–103. Springer, 2018b.

Hoffman, J., Roberts, D. A., and Yaida, S. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

Huang, H., Huang, Q., and Krahenbuhl, P. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 590–605, 2018.

Huang, L., Liu, L., Zhu, F., Wan, D., Yuan, Z., Li, B., and Shao, L. Controllable orthogonalization in training dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6429–6438, 2020.

Ilyas, A., Santurkar, S., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Koltchinskii, V. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Leen, T. K. From data distributions to regularization in invariant learning. *Neural Computation*, 7(5):974–981, 1995.

LeJeune, D., Balestriero, R., Javadi, H., and Baraniuk, R. G. Implicit rugosity regularization via data augmentation. *arXiv preprint arXiv:1905.11639*, 2019.

Li, L., Zhong, Z., Li, B., and Xie, T. Robustra: Training provable robust neural networks over reference adversarial space. In *IJCAI*, pp. 4711–4717, 2019.

Li, L., Qi, X., Xie, T., and Li, B. Sok: Certified robustness for deep neural networks. *arXiv*, abs/2009.04131, 2020.

Li, L., Weber, M., Xu, X., Rimanic, L., Kailkhura, B., Xie, T., Zhang, C., and Li, B. Tss: Transformation-specific smoothing for robustness certification. In *ACM Conference on Computer and Communications Security (CCS 2021)*, 2021.

Liang, K., Zhang, J. Y., Wang, B., Yang, Z., Koyejo, O., and Li, B. Uncovering the connections between adversarial transferability and knowledge transferability. *ICML*, 2020.

Lyle, C., Kwiatkowksa, M., and Gal, Y. An analysis of the effect of invariance on generalization in neural networks. In *International conference on machine learning Workshop on Understanding and Improving Generalization in Deep Learning*, volume 1, 2019.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*, 2009.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Rosenfeld, E., Ravikumar, P., and Risteski, A. An online learning approach to interpolation and extrapolation in domain generalization. *arXiv preprint arXiv:2102.13128*, 2021.

Roth, K., Kilcher, Y., and Hofmann, T. Adversarial training is a form of data-dependent operator norm regularization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

Utrera, F., Kravitz, E., Erichson, N. B., Khanna, R., and Mahoney, M. W. Adversarially-trained deep nets transfer better: Illustration on image classification. In *International Conference on Learning Representations*, 2020.

van der Wilk, M., Bauer, M., John, S., and Hensman, J. Learning invariances using the marginal likelihood. *arXiv preprint arXiv:1808.05563*, 2018.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pp. 1–6. IEEE, 2016.

Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., and Song, D. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 217–234, 2018.

Yang, Z., Li, L., Xu, X., Zuo, S., Chen, Q., Zhou, P., Rubinstein, B. I. P., Zhang, C., and Li, B. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. In *Neural Information Processing Systems (NeurIPS 2021)*, 2021.

You, K., Wang, X., Long, M., and Jordan, M. Towards accurate model selection in deep unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 7124–7133. PMLR, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.