Toward Learning Human-aligned Cross-domain Robust Models by Countering Misaligned Features

Abstract

Machine learning has demonstrated remarkable prediction accuracy over *i.i.d* data, but the accuracy often drops when tested with data from another distribution. In this paper, we aim to offer another view of this problem in a perspective assuming the reason behind this accuracy drop is the reliance of models on the features that are not aligned well with how a data annotator considers similar across these two datasets. We refer to these features as misaligned features. We extend the conventional generalization error bound to a new one for this setup with the knowledge of how the misaligned features are associated with the label. Our analysis offers a set of techniques for this problem, and these techniques are naturally linked to many previous methods in robust machine learning literature. We also compared the empirical strength of these methods demonstrated the performance when these previous techniques are combined.

1 INTRODUCTION

Machine learning, especially deep neural networks, has demonstrated remarkable empirical successes over various applications. The models even occasionally achieved results beyond human-level performances over benchmark datasets [*e.g.,* He et al., 2015]. However, whether it is desired for a model to outsmart human on benchmarks remains an open discussion in recent years: indeed, a model can create more application opportunities when it surpasses human-level performances, but the community also notices that the performance gain is sometimes due to model's exploitation of the features meaningless to a human, which may lead to unexpected performance drops when the models are tested with other datasets in practice that a human considers similar to the benchmark [Christian, 2020].

One of the most famous examples of the model's exploitation of non-human-aligned features is probably the usage of snow background in "husky vs. wolf" image classification [Ribeiro et al., 2016]. Briefly, when the model is trained to classify "husky vs wolf," it notices that wolf images usually have a snow background and directly learns to use the background features. This example is only one of many similar discussions concerning that the models are using features considered futile by humans [*e.g.,* Wang et al., 2019a, Bahng et al., 2019, Sun et al., 2019], and, sometimes, the features used are not even perceptible to a human [Geirhos et al., 2019, Ilyas et al., 2019, Wang et al., 2020a, Hermann and Kornblith, 2020]. The usage of these features may lead to a misalignment between the human and the models' understanding of the data, thus leading to a potential performance drop when the models' are applied to other data that a human considers similar.

We illustrate this challenge with a toy example in Figure 1, where the model is trained on the source domain data to classify triangle vs. circle and tested on the target domain data with a different marginal distribution. However, the color coincides with the shape on the source domain. As a result, the model may learn either the shape function or the color function. The color function will not classify the target domain data correctly while the shape function can, but the empirical risk minimizer (ERM) cannot differentiate them and may learn either the shape or the color, leading to potentially degraded performances during the test. As one may expect, whether shape or color is considered human-aligned is subjective depending on the task or the data and, in general, irrelevant to the statistical nature of the problem. Therefore, our remaining analysis will depend on such knowledge.

In this paper, we aim to formalize the above challenge to study the learning of human-aligned models. In particular, we derive a new generalization error bound when a model is trained on one distribution but tested on another distribution that a human considers similar. As discussed previously, one potential challenge for this scenario is that the model may

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

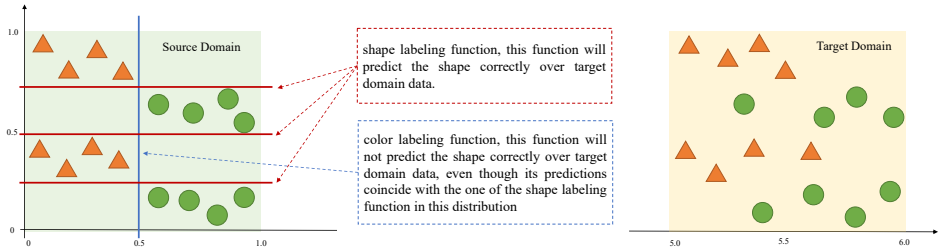

Figure 1: An illustration of the main problem focused in this paper.

learn to use some features, which we refer to as *misaligned features*, that a human considers irrelevant. Corresponding to this challenge, our analysis will be built upon the knowledge of how misaligned features are associated with the label.

2 RELATED WORK

There is a recent proliferation of methods aiming to learn robust models by enforcing the models to disregard certain features. We consider these works direct precedents of our discussion because these features are usually defined when comparing the model's performances to a human's. For example, the texture or background of images is probably the most often discussed misaligned feature for image classification. Below, we briefly discuss these works in two main strategies.

Data Augmentation With the knowledge of the misaligned features, the most effective solution is probably to augment the data by perturbing these misaligned features. Some recent examples of the perturbations used to train robust models include style transfer of images [Geirhos et al., 2019], naturalistic augmentation (color distortion, noise, and blur) of images [Hermann and Kornblith, 2020], other naturalistic augmentations (texture, rotation, contrast) of images [Wang et al., 2020b], interpolation of images [Hendrycks et al., 2019a], syntactic transformations of sentences [Mahabadi et al., 2020], and across data domain [Shankar et al., 2018, Lee et al., 2021].

Further, as recent studies suggest that one reason for the adversarial vulnerability of image classification [Szegedy et al., 2013, Goodfellow et al., 2015] is the existence of imperceptible features correlated with the label [Ilyas et al., 2019, Wang et al., 2020a], improving adversarial robustness may also be about countering the model's tendency toward learning these features. Currently, one of the most widely accepted methods to improve adversarial robustness is to augment the data along the training process to maximize the training loss by perturbing these features within predefined robustness constraints (*e.g.*, within ℓ_p norm ball) [Madry et al., 2018]. While this augmentation strategy is widely referred to as adversarial training, for the convenience of our later discussion, we refer to it as the worst-case data augmentation, following the naming conventions of [Fawzi et al., 2016].

Regularizing Hypothesis Space Another thread is to introduce inductive bias (*i.e.*, to regularize the hypothesis space) to force the model to discard misaligned features. To achieve this goal, one usually needs to first construct a side component to inform the main model about the misaligned features, and then to regularize the main model according to the side component. The construction of this side component usually relies on prior knowledge of what the misaligned features are. Then, methods can be built accordingly to counter the features such as the texture of images [Wang et al., 2019b, Bahng et al., 2019], the local patch of images [Wang et al., 2019a], label-associated keywords [He et al., 2019], label-associated text fragments [Mahabadi et al., 2020], and general easy-to-learn patterns of data [Nam et al., 2020].

In a broader scope, following the argument that one of the main challenges of domain adaptation is to counter the model's tendency in learning domain-specific features [*e.g.*, Ganin et al., 2016, Li et al., 2018], some methods contributing to domain adaption may have also progressed along the line of our interest. The most famous example is probably the domain adversarial neural network (DANN) [Ganin et al., 2016]. Inspired by the theory of domain adaptation [Ben-David et al., 2010], DANN trains the cross-domain generalizable neural network with the help of a side component specializing in classifying samples' domains. The subtle difference between this work and the ones mentioned previously is that this side component is not constructed with a special inductive bias but built as a simple network learning to classify domains with auxiliary annotations (domain IDs). DANN also inspires a family of methods forcing the model to learn auxiliary-annotation-invariant representations with a side component such as [Ghifary et al., 2016, Rozantsev et al., 2018, Motiian et al., 2017, Li et al., 2018, Carlucci et al., 2018].

Relation to Previous Works The above methods solve the same human-aligned learning problems with two different perspectives, but we notice the same central theme of forcing the models to *not* learn something according to the

prior knowledge of the data or the task. Although this central theme has been noticed by prior works such as [Wang et al., 2019b, Bahng et al., 2019, Mahabadi et al., 2020], we notice a lack of formal analysis from a task-agnostic viewpoint. Therefore, we continue to investigate whether we can contribute a principled understanding of this central theme, which serves as a connection of these methods and, potentially, a guideline for developing future methods. Also, we notice that many works along the domain adaptation development have rigorous statistical analysis [Ben-David et al., 2007, 2010, Mansour et al., 2009, Germain et al., 2016, Zhang et al., 2019, Dhouib et al., 2020], and these analyses mostly focus on the alignment of the distributions. Our study will complement these works by investigating through the perspective of misaligned features, and advantages and the disadvantages of our perspective will also be discussed.

3 GENERALIZATION UNDERSTANDING OF HUMAN-ALIGNED ROBUST MODELS

Roadmap We study the generalization error bound of human-aligned robust model in this section. We will first set up the problem of studying the generalization of the model across two distributions, whose difference mainly lies in the fact that one distribution has another labelling function (namely, the misaligned labelling function) in addition to the one that is shared across both of these distributions (**A2**). Then, to help quantify the error bound, we need to define the active set (features used by the function) ($\mathcal{A}(f, \mathbf{x})$ in (3)), the difference between the two functions ($d(\theta, f, \mathbf{x})$ in (4)), and an additional term to quantify whether the model learns the function if the model can map the sample correctly ($r(\theta, \mathcal{A}(f, \mathbf{x}))$ in (5)). With these terms defined, we will show a formal result on the generalization error bound, which depends on how many training samples are predicted correctly when the model learns the mis-aligned samples in addition to the standard terms.

3.1 NOTATIONS & BACKGROUND

We consider a binary classification problem from feature space $\mathcal{X} \in \mathbb{R}^p$ to label space $\mathcal{Y} \in \{0, 1\}$. The distribution over $\mathcal{X}$ is denoted as $\mathbf{P}$. A *labeling function* $f : \mathcal{X} \to \mathcal{Y}$ is a function that maps the feature $\mathbf{x}$ to its label $\mathbf{y}$. A *hypothesis* or *model* $\theta : \mathcal{X} \to \mathcal{Y}$ is also a function that maps the feature to the label. The difference in naming is only because we want to differentiate whether the function is a natural property of the space or distribution (thus called a labeling function) or a function to estimate (thus called a hypothesis or model). The hypothesis space is denoted as Θ. We use dom to denote the domain (input space) of a function, thus $\text{dom}(\theta) = \mathcal{X}$.

This work studies the generalization error across two distributions, namely source and target distribution, denoted as $\mathbf{P}_s$ and $\mathbf{P}_t$, respectively. We are only interested when these two distributions are, considered by a human, similar but different: being similar means there exists a *human-aligned labeling function*, f_h, that maps any $\mathbf{x} \in \mathcal{X}$ to its label (thus the label $\mathbf{y} := f_h(\mathbf{x})$); being different means there exists a *misaligned labeling function*, f_m, that for any $\mathbf{x} \sim \mathbf{P}_s$, $f_m(\mathbf{x}) = f_h(\mathbf{x})$. This "similar but different" property will be reiterated as an assumption (**A2**) later. We use $(\mathbf{x}, \mathbf{y})$ to denote a sample, and use $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}}$ to denote a finite dataset if the features are drawn from $\mathbf{P}$. We use $\epsilon_{\mathbf{P}}(\theta)$ to denote the expected risk of θ over distribution $\mathbf{P}$, and use $\widehat{\cdot}$ to denote the estimation of the term $\cdot$ (*e.g.*, the empirical risk is $\widehat{\epsilon}_{\mathbf{P}}(\widehat{\theta})$). We use $l(\cdot, \cdot)$ to denote a generic loss function.

For a dataset $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}}$, if we train a model with

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{(\mathbf{x}, \mathbf{y}) \in (\mathbf{X}, \mathbf{Y})_{\mathbf{P}}} l(\theta(\mathbf{x}), \mathbf{y}), \tag{1}$$

previous generalization study suggests that we can expect the error rate to be bounded as

$$\epsilon_{\mathbf{P}}(\widehat{\theta}) \leq \widehat{\epsilon}_{\mathbf{P}}(\widehat{\theta}) + \phi(|\Theta|, n, \delta), \tag{2}$$

where $\epsilon_{\mathbf{P}}(\widehat{\theta})$ and $\widehat{\epsilon}_{\mathbf{P}}(\widehat{\theta})$ respectively are

$$\epsilon_{\mathbf{P}}(\widehat{\theta}) = \mathbb{E}_{\mathbf{x} \sim \mathbf{P}} |\widehat{\theta}(\mathbf{x}) - \mathbf{y}| = \mathbb{E}_{\mathbf{x} \sim \mathbf{P}} |\widehat{\theta}(\mathbf{x}) - f_h(\mathbf{x})|$$

and

$$\widehat{\epsilon}_{\mathbf{P}}(\widehat{\theta}) = \frac{1}{n} \sum_{(\mathbf{x}, \mathbf{y}) \in (\mathbf{X}, \mathbf{Y})_{\mathbf{P}}} |\widehat{\theta}(\mathbf{x}) - \mathbf{y}|,$$

and $\phi(|\Theta|, n, \delta)$ is a function of hypothesis space $|\Theta|$, number of samples n, and δ accounts for the probability when the bound holds. This paper expands the discussion with this generic form that can relate to several discussions, each with its own assumptions. We refer to these assumptions as **A1**.

A1: basic assumptions needed to derived (2), for example,

- when **A1** is "Θ is finite, $l(\cdot, \cdot)$ is a zero-one loss, samples are *i.i.d*", $\phi(|\Theta|, n, \delta) = \sqrt{(\log(|\Theta|) + \log(1/\delta))/2n}$
- when **A1** is "samples are *i.i.d*", $\phi(|\Theta|, n, \delta) = 2\mathcal{R}(\mathcal{L}) + \sqrt{(\log 1/\delta)/2n}$, where $\mathcal{R}(\mathcal{L})$ stands for Rademacher complexity and $\mathcal{L} = \{l_\theta \,|\, \theta \in \Theta\}$, where l_θ is the loss function corresponding to θ.

For more information or more concrete examples of the generic term, we refer interested readers to relevant textbooks such as [Bousquet et al., 2003].

3.2 GENERALIZATION ERROR BOUND OF HUMAN-ALIGNED ROBUST MODELS

Formally, we state the challenge of our human-aligned robust learning problem as the assumption:

A2: **Existence of Misaligned Features:** For any $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} := f_h(\mathbf{x})$. We also have a f_m that is different from f_h, and for $\mathbf{x} \sim \mathbf{P}_s$, $f_h(\mathbf{x}) = f_m(\mathbf{x})$.

Thus, the existence of f_m is a key challenge for the small empirical risk over $\mathbf{P}_s$ to be generalized to $\mathbf{P}_t$, because θ that learns either f_h or f_m will lead to small source error, but only θ that learns f_h will lead to small target error. Note that f_m may not exist for an arbitrary $\mathbf{P}_s$. In other words, **A2** can be interpreted as a regulation to $\mathbf{P}_s$ so that f_m, while being different from f_h, exists for any $\mathbf{x} \sim \mathbf{P}_s$.

In this problem, f_m and f_h are not the same despite $f_m(\mathbf{x}) = f_h(\mathbf{x})$ for any $\mathbf{x} \sim \mathbf{P}_s$, and we consider the differences lie in the features they use. To describe this difference, we introduce the notation $\mathcal{A}(\cdot, \cdot)$, which denotes a set parametrized by the labeling function and the sample, to describe the *active set* of features used by the labeling function. By *active set*, we refer to the minimum set of features that a labeling function requires to map a sample to its label. Formally, we define

$$\begin{aligned} &\mathcal{A}(f, \mathbf{x}) = \{i | \widehat{\mathbf{z}}_i = \mathbf{x}_i\}, \quad \text{where,} \\ &\widehat{\mathbf{z}} = \underset{\mathbf{z} \in \text{dom}(f), f(\mathbf{z}) = f(\mathbf{x})}{\arg\min} |\{i | \mathbf{z}_i = \mathbf{x}_i\}|, \end{aligned} \tag{3}$$

and $|\cdot|$ measures the cardinality. Intuitively, $\mathcal{A}(f, \mathbf{x})$ indexes the features f uses to predict $\mathbf{x}$. Although $f_m(\mathbf{x}) = f_h(\mathbf{x})$, $\mathcal{A}(f_m, \mathbf{x})$ and $\mathcal{A}(f_h, \mathbf{x})$ can be different. $\mathcal{A}(f_m, \mathbf{x})$ is the *misaligned features* following our definition.

Further, we define a new function difference given a sample as

$$d(\theta, f, \mathbf{x}) = \max_{\mathbf{z} \in \text{dom}(f): \mathbf{z}_{\mathcal{A}(f,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f,\mathbf{x})}} |\theta(\mathbf{z}) - f(\mathbf{z})|, \tag{4}$$

where $\mathbf{x}_{\mathcal{A}(f,\mathbf{x})}$ denotes the features of $\mathbf{x}$ indexed by $\mathcal{A}(f, \mathbf{x})$. In other words, the distance describes: given a sample $\mathbf{x}$, the maximum disagreement of the two functions θ and f for all the other data $\mathbf{z} \in \mathcal{X}$ with a constraint that the features indexed by $\mathcal{A}(f, \mathbf{x})$ are the same as those of $\mathbf{x}$. Notice that this difference is not symmetric, as the active set is determined by the second function. By definition, we have $d(\theta, f, \mathbf{x}) \geq |\theta(\mathbf{x}) - f(\mathbf{x})|$.

Also, please notice that when we use expressions such as $\mathbf{z}_{\mathcal{A}(f,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f,\mathbf{x})}$, we believe the $\mathcal{A}(f, \mathbf{x})$ in both LHS and RHS should refer to the same set of indices of features. Under this premise of the notation, whether (3) has a unique solution or not will not affect our main conclusion.

In addition, one may notice the connection between $\mathcal{A}(f, \mathbf{x})$ and the minimum sufficient explanation discussed previously [*e.g.*, Camburu et al., 2020, Yoon et al., 2019, Carter et al., 2019, Ribeiro et al., 2018]. While $\mathcal{A}(f, \mathbf{x})$ is conceptually the same as the minimum set of features for a model to predict, we define it mathematically different from previous works.

To continue, we introduce the following assumption:

A3: **Realized Hypothesis:** Given a large enough hypothesis space Θ, for any sample $(\mathbf{x}, \mathbf{y})$, for any $\theta \in \Theta$, which is not a constant mapping, if $\theta(\mathbf{x}) = \mathbf{y}$, then $d(\theta, f_h, \mathbf{x}) d(\theta, f_m, \mathbf{x}) = 0$

Intuitively, **A3** assumes θ at least learns one labeling function for the sample $\mathbf{x}$ if θ can map the $\mathbf{x}$ correctly.

Finally, to describe how θ depends on the active set of f, we introduce the term

$$r(\theta, \mathcal{A}(f, \mathbf{x})) = \max_{\mathbf{z}_{\mathcal{A}(f,\mathbf{x})} \in \text{dom}(f)_{\mathcal{A}(f,\mathbf{x})}} |\theta(\mathbf{z}) - \mathbf{y}|, \tag{5}$$

where $\mathbf{z}_{\mathcal{A}(f,\mathbf{x})} \in \text{dom}(f)_{\mathcal{A}(f,\mathbf{x})}$ denotes that the features of $\mathbf{z}$ indexed by $\mathcal{A}(f, \mathbf{x})$ are searched in the input space $\text{dom}(f)$. Notice that $r(\theta, \mathcal{A}(f, \mathbf{x})) = 1$ alone does not mean θ depends on the active set of f; it only means so when we also have $\theta(\mathbf{x}) = \mathbf{y}$ (see the formal discussion in Lemma A.1). In other words, $r(\theta, \mathcal{A}(f, \mathbf{x})) = 1$ alone may not have an intuitive meaning, but given $\theta(\mathbf{x}) = \mathbf{y}$, $r(\theta, \mathcal{A}(f, \mathbf{x})) = 1$ intuitively means θ learns f.

With all above, we can extend the conventional generalization error bound with a new term as follows:

Theorem 3.1. *With Assumptions **A1**-**A3**, $l(\cdot, \cdot)$ is a zero-one loss, with probability as least $1 - \delta$, we have*

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta) + \phi(|\Theta|, n, \delta) \tag{6}$$

where

$$c(\theta) = \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_m, \mathbf{x})).$$

$\mathbb{I}[\cdot]$ is a function that returns 1 if the condition $\cdot$ holds and 0 otherwise. As θ may learn f_m, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$ can no longer alone indicate $\epsilon_{\mathbf{P}_t}(\theta)$; thus, we introduce $c(\theta)$ to account for the discrepancy. Intuitively, $c(\theta)$ quantifies the samples that are correctly predicted, but only because the θ learns f_m for that sample. $c(\theta)$ critically depends on the knowledge of f_m.

3.3 IN COMPARISON TO THE VIEW OF DOMAIN ADAPTATION

We continue to compare Theorem 3.1 with understandings of domain adaptation. Conveniently, several domain adaptation analyses [Ben-David et al., 2007, 2010, Mansour et al., 2009, Germain et al., 2016, Zhang et al., 2019, Dhouib et al., 2020] can be sketched in the following form:

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta(\mathbf{P}_s, \mathbf{P}_t) + \lambda + \phi'(|\Theta|, n, \delta) \tag{7}$$

where $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$ quantifies the differences between the two distributions; λ describes the nature of the problem and

usually involves non-estimable terms about the problem or the distributions.

For example, Ben-David et al. [2010] formalized the difference as Θ-divergence, and described the corresponding empirical term as (with $\Theta\Delta\Theta$ denoting the set of disagreement between two hypotheses in Θ):

$$D_\Theta(\mathbf{P}_s, \mathbf{P}_t) = 1 - \min_{\theta \in \Theta\Delta\Theta} \Big(\frac{1}{n} \sum_{\mathbf{x}:\theta(\mathbf{x})=0} \mathbb{I}[\mathbf{x} \in (\mathbf{X}, \mathbf{Y})_{\mathbf{P}_s}] + \frac{1}{n} \sum_{\mathbf{x}:\theta(\mathbf{x})=1} \mathbb{I}[\mathbf{x} \in (\mathbf{X}, \mathbf{Y})_{\mathbf{P}_t}]\Big). \tag{8}$$

Also, Ben-David et al. [2010] formalized $\lambda = \epsilon_{\mathbf{P}_t}(\theta^\star) + \epsilon_{\mathbf{P}_s}(\theta^\star)$, where $\theta^\star = \arg\min_{\theta\in\Theta} \epsilon_{\mathbf{P}_t}(\theta) + \epsilon_{\mathbf{P}_s}(\theta)$,

In our discussion, as we assume the f_h applies to any $\mathbf{x} \in \mathcal{X}$ (according to **A2**), $\lambda = 0$ as long as the hypothesis space is large enough. Therefore, the comparison mainly lies in comparing $c(\theta)$ and $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$.

To compare them, we need an extra assumption:

A5: **Sufficiency of Training Samples** for the two finite datasets in the study, *i.e.*, $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}_s}$ and $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}_t}$, for any $\mathbf{x} \in (\mathbf{X}, \mathbf{Y})_{\mathbf{P}_t}$, there exists one or many $\mathbf{z} \in (\mathbf{X}, \mathbf{Y})_{\mathbf{P}_s}$ such that

$$\mathbf{x} \in \{\mathbf{x}' | \mathbf{x}' \in \mathcal{X} \text{ and } \mathbf{x}'_{\mathcal{A}(f_d,\mathbf{z})} = \mathbf{z}_{\mathcal{A}(f_d,\mathbf{z})}\} \tag{9}$$

A5 intuitively means the finite training dataset needs to be diverse enough to describe the concept that needs to be learned. For example, imagine building a classifier to classify mammals *vs.* fishes from the distribution of photos to that of sketches, we cannot expect the classifier to do anything good on dolphins if dolphins only appear in the test sketch dataset. **A5** intuitively regulates that if dolphins will appear in the test sketch dataset, they must also appear in the training dataset.

Now, in comparison to [Ben-David et al., 2010], we have

Theorem 3.2. *With Assumptions **A2**-**A5**, and if* $1 - f_d \in \Theta$, *we have*

$$c(\theta) \leq D_\Theta(\mathbf{P}_s, \mathbf{P}_t) + \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_m, \mathbf{x})) \tag{10}$$

where

$$c(\theta) = \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_m, \mathbf{x}))$$

and $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$ *is defined as in* (8).

The comparison involves an extra term, $q(\theta) := \frac{1}{n}\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_m, \mathbf{x}))$, which intuitively means that if θ learns f_m, how many samples θ can coincidentally predict correctly over the finite target set used to estimate $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$. For sanity check, if we replace $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}_t}$ with $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}_s}$, $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$ will be evaluated at 0 as it cannot differentiate two identical datasets, and $q(\theta)$ will be the same as $c(\theta)$. On the other hand, if no samples from $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}_t}$ can be mapped correctly with f_m (coincidentally), $q(\theta) = 0$ and $c(\theta)$ will be a lower bound of $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$.

The value of Theorem 3.2 lies in the fact that for an arbitrary target dataset $(\mathbf{X}, \mathbf{Y})_{\mathbf{P}_t}$, no samples out of which can be predicted correctly by learning f_m (a situation likely to occur for arbitrary datasets), $c(\theta)$ will always be a lower bound of $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$.

Further, when Assumption **A5** does not hold, we are unable to derive a clear relationship between $c(\theta)$ and $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$. The difference is mainly raised as a matter of fact that, intuitively, we are only interested in the problems that are "solvable" (**A5**, *i.e.*, hypothesis that used to reduce the test error in target distribution can be learned from the finite training samples) but "hard to solve" (**A2**, *i.e.*, another labeling function, namely f_m, exists for features sampled from the source distribution only), while $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$ estimates the divergence of two arbitrary distributions.

3.4 ESTIMATION OF $c(\theta)$

The estimation of $c(\theta)$ mainly involves two challenges: the requirement of the knowledge of f_m and the computational cost to search over the entire space $\mathcal{X}$.

The first challenge is unavoidable by definition because the human-aligned learning has to be built upon the prior knowledge of what labeling function a human considers similar (what f_h is) or its opposite (what f_m is). Fortunately, as discussed in Section 2, the methods are usually developed with prior knowledge of what the misaligned features are, suggesting that we may often directly have the knowledge of $\mathcal{A}(f_m, \mathbf{x})$.

The second challenge is about the computational cost to search, and the community has several techniques to help reduce the burden. For example, the search can be terminated once $r(\theta, \mathcal{A}(f_m, \mathbf{x}))$ is evaluated as 1 (*i.e.*, once we find a perturbation of misaligned features that alters the prediction). This procedure is similar to how adversarial attack [Goodfellow et al., 2015] is used to evaluate the robustness of models. To further reduce the computational cost, one can also generate some out-of-domain data by perturbing misaligned features beforehand and use these fixed data to test multiple models. Using fixed data to evaluate may not be as accurate as using a search process, but sometimes, it may be good enough to reveal some properties of the models

worth further attention [Jo and Bengio, 2017, Geirhos et al., 2019, Wang et al., 2020a].

4 METHODS TO LEARN HUMAN-ALIGNED ROBUST MODELS

We continue to study how our analytical results above can lead to practical methods to learn human-aligned robust models. We first show that our discussion can naturally connect to existing methods for robust machine learning discussed in Section 2. Further, as these methods mostly require some prior knowledge of the misaligned features, we continue to explore a new method that does not require so.

Theorem 3.1 suggests that training a human-aligned robust model amounts to training for small $c(\theta)$ and small empirical error (*i.e.*, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$).

4.1 WORST-CASE TRAINING

To simplify the notation, we define $\mathcal{Q}(\mathbf{x}) := \{\mathbf{x}_{\mathcal{A}(f_m,\mathbf{x})} \in \text{dom}(f_m)_{\mathcal{A}(f_m,\mathbf{x})}\}$. We can consider the upper bound of $c(\theta)$

$$\begin{aligned} c(\theta) \leq & \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} r(\theta, \mathcal{A}(f_m, \mathbf{x})) \\ = & \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \max_{\mathbf{z}\in\mathcal{Q}(\mathbf{x})} |\theta(\mathbf{z}) - \mathbf{y}|, \end{aligned} \tag{11}$$

which intuitively means that instead of $c(\theta)$ that studies only the correct predictions because θ learns f_m, now we study any predictions because θ learns f_m.

Further, as

$$|\theta(\mathbf{x}) - \mathbf{y}| \leq \max_{\mathbf{z}\in\mathcal{Q}(\mathbf{x})} |\theta(\mathbf{z}) - \mathbf{y}|,$$

a model with minimum (11) naturally means the model will have a minimum empirical loss. Therefore, we can train for a small (11), which likely leads to the model with a small empirical loss. Therefore, after we replace $|\theta(\mathbf{x}) - \mathbf{y}|$ with a generic loss term $\ell(\theta(\mathbf{x}), \mathbf{y})$, we can directly train a model with

$$\min_{\theta\in\Theta} \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \max_{\mathbf{z}\in\mathcal{Q}(\mathbf{x})} \ell(\theta(\mathbf{z}), \mathbf{y}) \tag{12}$$

to get a model with small $c(\theta)$ and small empirical error.

The above method is to augment the data by perturbing the misaligned features to maximize the training loss and solve the optimization problem with the augmented data. This method is the worst-case data augmentation method [Fawzi et al., 2016] we discussed previously, and is also closely connected to one of the most widely accepted methods for the adversarial robust problem, namely the adversarial training [Madry et al., 2018].

While the above result shows that a method for learning human-aligned robust models is in mathematical connection to the worst-case data augmentation, in practice, a general application of this method will require some additional assumptions. The detailed discussions of these are in the appendix.

We continue from the RHS of (11) to discuss another reformulation by reweighting sample losses for optimization, which leads to:

$$\frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \max_{\mathbf{z}\in\mathcal{Q}(\mathbf{x})} \lambda(\mathbf{z})|\theta(\mathbf{z}) - \mathbf{y}| \tag{13}$$

The conditions (assumptions) that we need for $c(\theta) \leq$ the LHS of (13) is discussed in the appendix. Now, we will continue with

$$c(\theta) \leq \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \max_{\mathbf{z}\in\mathcal{Q}(\mathbf{x})} \lambda(\mathbf{z})|\theta(\mathbf{z}) - \mathbf{y}| \tag{14}$$

When (14) holds, replacing $|\theta(\mathbf{z}) - \mathbf{y}|$ with a generic loss $\ell(\theta(\mathbf{z}), \mathbf{y})$ and minimizing it is another direction of learning robust models, which corresponds to distributionally robust optimization (DRO) [Ben-Tal et al., 2013, Duchi et al., 2021].

Further, depends on implementations of $\lambda(\mathbf{x})$, DRO has been implemented with different concrete solutions, sometimes with structural assumptions [Hu et al., 2018], such as

- Adversarially reweighted learning (ARL) [Lahoti et al., 2020] uses another model $\phi : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ to identify samples with misaligned features that cause high losses of model θ and defines

$$\lambda(\mathbf{x}) = 1 + |(\mathbf{X}, \mathbf{Y})| \cdot \frac{\phi(\mathbf{x})}{\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \phi(\mathbf{x})}$$

- Learning from failures (LFF) [Nam et al., 2020] also trains another model ϕ by amplifying its early-stage predictions and defines

$$\lambda(\mathbf{x}) = \frac{\ell(\phi(\mathbf{x}), \mathbf{y})}{\ell(\phi(\mathbf{x}), \mathbf{y}) + \ell(\theta(\mathbf{x}), \mathbf{y})} \tag{15}$$

- Group DRO [Sagawa* et al., 2020] assumes the availability of the structural partition of the samples, and defines the weight of samples at partition $\mathbf{g}$ as

$$\lambda(\mathbf{x}) = \frac{\exp(\ell(\theta(\mathbf{x}), \mathbf{y}))}{\sum_{(\mathbf{z},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{g}}} \exp(\ell(\theta(\mathbf{z}), \mathbf{y}))}, \tag{16}$$

  if $(\mathbf{x}, \mathbf{y}) \in (\mathbf{X}, \mathbf{Y})_{\mathbf{g}}$, samples of partition $\mathbf{g}$

These discussions are expanded in the appendix.

4.2 REGULARIZING THE HYPOTHESIS SPACE

Connecting our theory to the other main thread is little bit tricky, as we need to extend the model to an encoder/decoder structure, where we use e_θ and d_θ to denote the encoder and decoder respectively. Thus, by definition, we have $\theta(\mathbf{x}) = d_\theta(e_\theta(\mathbf{x}))$. Further, we define f'_m as the equivalent of f_m with the only difference is that f'_m operates on the representations $e_\theta(\mathbf{x})$. With the setup, optimizing the empirical loss and $c(\theta)$ leads to (details in the appendix):

$$\min_{d_\theta, e_\theta} \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \ell(d_\theta(e_\theta(\mathbf{x})), \mathbf{y}) - \ell(f'_m(e_\theta(\mathbf{x})), \mathbf{y}), \tag{17}$$

which is highly related to methods used to learn auxiliary-annotation-invariant representations, and the most popular example of these methods is probably DANN [Ganin et al., 2016].

Then, the question left is how to get f'_m. We can design a specific architecture given the prior knowledge of the data, then f'_m can be directly estimated through

$$\min_{f'_m} \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \ell(f'_m(e_\theta(\mathbf{x})), \mathbf{y}), \tag{18}$$

which connects to several methods in Section 2, such as [Wang et al., 2019a, Bahng et al., 2019]. Alternatively, we can estimate f'_m with additional annotations (*e.g.*, domain ids, batch ids *etc*), then we can estimate the model by (with $\mathbf{t}$ denoting the additional annotation)

$$\min_{f'_m} \frac{1}{n} \sum_{(\mathbf{x},\mathbf{t})\in(\mathbf{X},\mathbf{T})} \ell(f'_m(e_\theta(\mathbf{x})), \mathbf{t}), \tag{19}$$

which connects to methods in domain adaptation literature such as [Ganin et al., 2016, Li et al., 2018].

4.3 A NEW HEURISTIC: WORST-CASE TRAINING WITH REGULARIZED HYPOTHESIS SPACE

Our analysis showed that optimizing for small $c(\theta)$ naturally connects to one of the two mainstream families of methods used to train robust models in the literature, which naturally inspires us to invent a new method by combining these two directions. The intuition behind this design rationale is to incorporate the empirical strength of each of these methods together by directly combining the major components of these methods. z

Therefore, we introduce a new heuristic that combines the worst-case training (11) and the regularization method (17) and (19), for which, whether the samples are originally from $(\mathbf{X}, \mathbf{Y})$ or generated along the training will serve as the additional annotation $\mathbf{t}$.

Algorithm 1: worst-case training with regularized hypothesis space

Result: θ^I
Input: total iterations I, $(\mathbf{X}, \mathbf{Y})$;
initialize $\theta^{(0)}$, $i = 1$;
while $i \leq I$ **do**
  for *sample* $(\mathbf{x}, \mathbf{y})$ **do**
    assign additional label $\mathbf{t}_\mathbf{x} = 0$ for $\mathbf{x}$;
    sample $\mathbf{z} \in Q(\mathbf{x})$ that maximizes $\ell(\theta(\mathbf{x}), \mathbf{y})$;
    assign additional label $\mathbf{t}_\mathbf{z} = 1$ for $\mathbf{z}$;
    update f'_m with (19);
    update θ with (17);
    update θ with $\mathbf{z}$ with the equivalence of (17) as
     $\min_{d_\theta, e_\theta} \ell(d_\theta(e_\theta(\mathbf{z})), \mathbf{y}) - \ell(f'_m(e_\theta(\mathbf{z})), \mathbf{y})$,
  end
end

In particular, our heuristic is illustrated with Algorithm 1. In practice, we will also introduce a hyperparamter to balance the two losses in (17).

5 EXPERIMENTS

We will present our theory supporting experiments in the Appendix, and mainly discuss the empirical performance of our proposed heuristic here.

To test the performance of our newly introduced heuristic, we compare our methods on a fairly recent and strong baseline. In particular, we follow the setup of a direct precedent of our work [Bahng et al., 2019] to compare the models for a nine super-class ImageNet classification [Ilyas et al., 2019] with class-balanced strategies. Also, we follow [Bahng et al., 2019] to report standard accuracy, weighted accuracy, a scenario where samples with unusual texture are weighted more, and accuracy over ImageNet-A [Hendrycks et al., 2019b], a collection of failure cases for most ImageNet trained models. Additionally, we also report the performance over ImageNet-Sketch [Wang et al., 2019a], an independently collected ImageNet test set with only sketch images.

We test our method with the pipeline made available by [Bahng et al., 2019], and we compare with the vanilla network, and several methods that are designed to reduce the texture bias: including StylisedIN (SN) [Geirhos et al., 2019], LearnedMixin (LM) [Clark et al., 2019], RUBi [Cadene et al., 2019], and ReBias [Bahng et al., 2019], several other baselines proved effective in learning robust models, such as Mix-up [Zhang et al., 2017], Cutout [DeVries and Taylor, 2017], AugMix Hendrycks et al. [2019a], In addition, we compared our worst-case training method (WT), regularization method (Reg), and our introduced heuristic (WR). For our methods, we follow the observations in [Wang et al., 2020a], which suggests the relationship between frequency-

	Vanilla	SN	LM	RUBi	ReBias	Mixup	Cutout	AugMix	WT	Reg	WR
Standard Acc.	90.80	88.40	67.90	90.50	91.90	92.50	91.20	92.90	92.50	93.10	**93.30**
Weighted Acc.	88.80	86.60	65.90	88.60	90.50	91.20	90.30	91.70	91.30	**92.20**	92.00
ImageNet-A	24.90	24.60	18.80	27.70	29.60	29.10	27.30	**31.50**	28.50	30.00	29.60
ImageNet-Sketch	41.10	40.50	36.80	42.30	41.80	40.60	38.70	41.40	43.00	42.50	**43.20**
average	61.40	60.03	47.35	62.28	63.45	63.35	61.88	64.38	63.83	64.45	**64.53**

Table 1: Results comparison on nine super-class ImageNet classification.

based perturbation and the model's performance, and design the augmentation of frequency-based perturbation with different radii.

We report the results in Table 1. Our results suggest that, while the augmentation method we used is much simpler than the ones used in AugMix, our empirical results are fairly strong in comparison. With simple perturbation inspired from [Wang et al., 2020a], our new heuristic outperforms other methods in average on these four different test scenarios.

6 DISCUSSION

Before we conclude, we would like to devote a section to discuss several topics more broadly related to this paper. We hope the discussion can also help define the scope of our contributions.

Human-aligned machine learning may not be solvable in general without prior knowledge. Following our notations in this paper, for any two functions f_1 and f_2, it is human, instead of any statistical properties, that decides whether $f_h = f_1$ or $f_h = f_2$. This remark is a restatement of our motivating example in Figure 1. Our proposed method forgoes the requirement of prior knowledge and is validated empirically on certain benchmark datasets.

Do all the model's understandings of the data have to be aligned with a human's? Probably no. As we have discussed in the preceding sections, we agree that there are also scenarios where it is beneficial for models' perception to outperform a human's. For example, we may expect the models to outperform the human vision system when applied to make a scientific discovery at a molecule level. This paper investigates these questions for the scenarios where the alignment is essential.

In practice, there is probably more than one source of misaligned features. We aim to contribute a principled understanding of the problem, starting with its basic form. The extension of our analysis to multiple sources of misaligned features is considered a future direction.

Other related works There is also a proliferation of works that aim to improve the robustness of machine learning methods from a data perspective, such as the methods developed to counter spurious correlations [Vigen, 2015], confounding factors [McDonald, 2014], or dataset bias [Torralba and Efros, 2011]. We believe how our analysis is statistically connected to these topics is also an interesting direction to pursue in the future. Further, there is also an active line of research aiming to align the human and models' perception of data by directly studying how humans process the images [Kubilius et al., 2019, Marblestone et al., 2016, Nayebi and Ganguli, 2017, Lindsay and Miller, 2018, Li et al., 2019, Dapello et al., 2020].

In addition, discussion of how human annotation will help the models to generalize in non-trivial test scenarios has been explored previously. For example, [Ross et al., 2017] built expert annotation into the model to regularize the explanation of the models to counter the model's tendency in learning misaligned features. The study has been extended with multiple follow-ups to introduce human-annotation into the interpretation of the models [Schramowski et al., 2020, Teso and Kersting, 2019, Lertvittayakumjorn et al., 2020], and shows that the human's knowledge will help model's learning the concepts that can generalize in non-i.i.d scenarios.

7 CONCLUSION

In this paper, we built upon the premise that it is important to align the models' understanding of data with that of a human and studied the generalization properties of a model for the goal of the alignment between the human and the model. We extended the widely-accepted generalization error bound with an additional term that accounts for the differences between the human and the model, and this new term relies on how the misaligned features are associated with the label. According to our analysis, optimizing for small empirical loss and small this term will lead to a model that is better aligned to humans. Thus, our analysis naturally offers a set of methods to this problem. Interestingly, the methods our analysis leads to are closely connected to the established methods in multiple topics regarding robust machine learning. Finally, by noticing our analysis can link to two mainstream families of methods discussed in the context of learning robust models, we propose a new heuristic of combining these two methods. In a fairly advanced experiment, we demonstrate the empirical strength of our new method.

References

Hyojin Bahng et al. Learning de-biased representations with biased representations. *arXiv:1910.02806*, 2019.

Shai Ben-David et al. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137–144, 2007.

Shai Ben-David et al. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Aharon Ben-Tal et al. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Olivier Bousquet et al. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

Remi Cadene et al. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, pages 841–852, 2019.

Oana-Maria Camburu et al. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets. *arXiv:2009.11023*, 2020.

Fabio M Carlucci et al. Agnostic domain generalization. *arXiv:1808.01102*, 2018.

Brandon Carter et al. What made you do this? understanding black-box decisions with sufficient input subsets. In *The 22nd AISTATS*, pages 567–576. PMLR, 2019.

Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company, 2020.

Christopher Clark et al. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv:1909.03683*, 2019.

Joel Dapello et al. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In *NeurIPS 2020*, 2020.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv:1708.04552*, 2017.

Sofien Dhouib et al. Margin-aware adversarial domain adaptation with optimal transport. In *Thirty-seventh ICML*, 2020.

John C Duchi et al. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

Alhussein Fawzi et al. Adaptive data augmentation for image classification. In *ICIP 2016*, pages 3688–3692. IEEE, 2016.

Yaroslav Ganin et al. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

Robert Geirhos et al. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Pascal Germain et al. A new pac-bayesian perspective on domain adaptation. In *ICML, 2016*, volume 48, pages 859–868, 2016.

Muhammad Ghifary et al. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, pages 597–613. Springer, 2016.

Ian J Goodfellow et al. Explaining and harnessing adversarial examples (2014). In *ICLR*, 2015.

He He et al. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv:1908.10763*, 2019.

Kaiming He et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015.

Dan Hendrycks et al. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2019a.

Dan Hendrycks et al. Natural adversarial examples. *arXiv:1907.07174*, 2019b.

Katherine Hermann and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *NeurIPS*, 33, 2020.

Weihua Hu et al. Does distributionally robust supervised learning give robust classifiers? In *ICML*, pages 2029–2037. PMLR, 2018.

Andrew Ilyas et al. Adversarial examples are not bugs, they are features. In *NeurIPS*, pages 125–136, 2019.

Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *CoRR*, abs/1711.11561, 2017. URL http://arxiv.org/abs/1711.11561.

Jonas Kubilius et al. Brain-like object recognition with high-performing shallow recurrent anns. In *NeurIPS 2019*, pages 12785–12796, 2019.

Preethi Lahoti et al. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, volume 33, pages 728–740, 2020.

Saehyung Lee et al. Removing undesirable feature contributions using out-of-distribution data. *arXiv:2101.06639*, 2021.

Piyawat Lertvittayakumjorn et al. Find: human-in-the-loop debugging deep text classifiers. *arXiv:2010.04987*, 2020.

Haoliang Li et al. Domain generalization with adversarial feature learning. In *CVPR*, 2018.

Zhe Li et al. Learning from brains how to regularize machines. In *NeurIPS 2019*, pages 9525–9535, 2019.

Grace W Lindsay and Kenneth D Miller. How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, 7:e38105, 2018.

Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In *ICLR 2018*, 2018.

Rabeeh Karimi Mahabadi et al. End-to-end bias mitigation by modelling biases in corpora. In *ACL 2020*, pages 8706–8716. ACL, 2020.

Yishay Mansour et al. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

Adam H. Marblestone et al. Toward an integration of deep learning and neuroscience. *Frontiers Comput. Neurosci.*, 10:94, 2016.

JH McDonald. Confounding variables. *Handbook of biological statistics, 3rd edn. Sparky House Publishing, Baltimore*, pages 24–28, 2014.

Saeid Motiian et al. Unified deep supervised domain adaptation and generalization. In *ICCV*, volume 2, page 3, 2017.

Junhyun Nam et al. Learning from failure: Training debiased classifier from biased classifier. In *NeurIPS*, 2020.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NeurIPS*, volume 29. Curran Associates, Inc., 2016.

Aran Nayebi and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. *arXiv:1703.09202*, 2017.

Marco Tulio Ribeiro et al. " why should i trust you?" explaining the predictions of any classifier. In *KDD 2017=6*, pages 1135–1144, 2016.

Marco Tulio Ribeiro et al. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 32, 2018.

Andrew Slavin Ross et al. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv:1703.03717*, 2017.

Artem Rozantsev et al. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018.

Shiori Sagawa*, Pang Wei Koh*, et al. Distributionally robust neural networks. In *ICLR*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Patrick Schramowski et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8): 476–486, 2020.

Shiv Shankar et al. Generalizing across domains via cross-gradient training. In *6th ICLR, ICLR 2018*, 2018.

Tony Sun et al. Mitigating gender bias in natural language processing: Literature review. In *ACL 2019*, pages 1630–1640. ACL, 2019.

Christian Szegedy et al. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

Tyler Vigen. *Spurious correlations*. Hachette books, 2015.

Haohan Wang et al. Learning robust global representations by penalizing local predictive power. In *NeurIPS 2019*, pages 10506–10518, 2019a.

Haohan Wang et al. Learning robust representations by projecting superficial statistics out. In *ICLR 2019*. OpenReview.net, 2019b.

Haohan Wang et al. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, pages 8684–8694, 2020a.

Haohan Wang et al. Squared ℓ_2 norm as consistency loss for leveraging augmented data to learn robust and invariant representations. *arXiv:2011.13052*, 2020b.

Jinsung Yoon et al. INVASE: Instance-wise variable selection using neural networks. In *ICLR*, 2019.

Hongyi Zhang et al. mixup: Beyond empirical risk minimization. *arXiv:1710.09412*, 2017.

Yuchen Zhang et al. Bridging theory and algorithm for domain adaptation. In *ICML, 2019*, volume 97, pages 7404–7413. PMLR, 2019.