Thesis for The Degree of Licentiate of Engineering

Towards practical and provable domain adaptation

Understanding the roles of data and assumptions

Adam Breitholtz

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Towards practical and provable domain adaptation
Understanding the roles of data and assumptions

Adam Breitholtz

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2023.

Till min familj.

Towards practical and provable domain adaptation
Understanding the roles of data and assumptions

Adam Breitholtz

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

One of the most central questions in statistical modeling is how well a model will generalize. Absent strong assumptions we find that this question is difficult to answer in a meaningful way. In this work we seek to increase our understanding of the domain adaptation setting through two different lenses. First, we investigate whether tractably computable and tight generalization bounds on the performance of neural network classifiers exist in the current literature. The tightest bounds we find use a portion of the input data to tighten the gap between measured performance and the calculated bound. We present evaluations of four bounds using this tightening method on classifiers applied to image classification tasks: Two bounds from the literature in addition to two of our own construction. Further, we find that for situations lacking domain overlap, the existing literature lacks the tools to achieve tight, tractably computable bounds for the neural network models which we use. We conclude that a new approach might be needed. In the second part we therefore consider a setting where we change our underlying assumptions to ones which might be more plausible. This setting, based on learning using privileged information, is shown to result in consistent learning. We also show empirical gains over comparable methods when our assumptions are likely to hold, both in terms of performance and sample efficiency. In summary, the work set out herein has been a first step towards a better understanding of domain adaptation and how using data and new assumptions can help us further our knowledge about this topic.

Keywords

Domain adaptation, Generalization, PAC-Bayes, Privileged information

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **A. Breitholtz**, F. D. Johansson, *Practicality of generalization guarantees for unsupervised domain adaptation with neural networks Transactions on Machine Learning Research (October, 2022).*

[**Paper II**] **A. Breitholtz**, A. Matsson, F. D. Johansson, *Unsupervised domain adaptation by learning using privileged information Submitted, under review.*

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] **A. Breitholtz**, F. D. Johansson, *Data Dependent Priors for Domain Adaptation Bounds*
The AAAI-22 Workshop on Engineering Dependable and Secure Machine Learning Systems (March, 2022).

Acknowledgment

I would like to start by thanking my supervisor Fredrik Johansson for his kind guidance and support. The work presented herein would not have been possible if not for your knowledge and drive. I also thank my co-supervisor, Devdatt Dubhashi, and my examiner, Dag Wedelin for their encouragement and kind feedback.

I would also like to express my gratitude towards the people of the DSAI division for making my workplace a more fun and exciting place. In particular, I want to thank my PhD colleagues: Arman, Alexander, Emilio, Filip, Daniel, Juan, Tobias, Christopher, Markus, Firooz, Mehrdad, David, Hampus, Emil, Fazeleh, Tobias, Niklas, Hanna, Linus and Hannes. A special thank you to the office buddies in 5453: Anton, Lena, Newton and Lovisa. I also want to thank my friend Samuel for his helpful insights and willingness to discuss these complex topics with me. Finally, I want to thank my family. Thank you Fredrik and Lena for being such supportive and loving parents and thank you Agnes for being the best sister a brother could ask for.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Contents

Part I

Summary

Chapter 1

Introduction

The point of fitting most statistical models is actually quite a simple one. We want the models to be sufficiently good at solving a specific task so we can deploy them in real-world settings to do this. The underlying goal is to train these models so that they can perform both during training and in deployment. We call this concept *generalization* and it has been extensively studied for all different kinds of mathematical models which have been produced.

In this work, we will consider a sub-problem of generalization called *domain adaptation* (DA). In DA we are considering the case where we are applying our model on a specific task but the underlying data distribution changes between training and deployment. The distribution over the input features and corresponding labels is called a domain, in application this could be represented by a change in location where the data is collected. For example, we collect weather data in southern Sweden to predict precipitation but apply the model in the north, where the underlying data distribution might be different.

DA is often a fairly realistic setting which we can find instantiated in myriad real-world applications. An example of this is in the healthcare setting when we are trying to classify pathologies from chest X-ray images. We could train a model to do this from data collected at Hospital A. We would then want to apply this model at Hospital B, however, the patient cohort at Hospital B can be substantially different than the one at A. In this setting it is commonplace that the labels corresponding to some features might be hard or even impossible to access, e.g. predicting outcomes which have yet to occur. For example, we probably do not know if a patient will develop lung cancer one year from when features are collected. Therefore we will allow ourselves to have access to features (X-ray images) from hospital B but not the corresponding labels. This specific setting is called *Unsupervised* domain adaptation (UDA).

To solve problems in the UDA setting a number of different approaches and methods have been proposed. Specifically, how to use the available data from the target is often considered. The approaches here are manifold; learn a representation which seeks to minimise some distance metric between the domains (Long et al., 2015), learn an adversarial classifier which tries to tell the two domains apart (Ganin et al., 2016), re-weight your data to more accurately

fit the target (Shimodaira, 2000), predict and use pseudo-labels on the target data using a model trained on source data (Saito et al., 2017) etc. However, in spite of all these advances there still remains a gap in performance compared to models that has been given access to target labels. This gap is so far not adequately explained by the theory. In addition, these methods all rely on assumptions which may or may not be realistic. These two gaps will be our main concern in this thesis.

In this thesis we will investigate ways in which we might achieve successful domain adaptation while still having useful guarantees on performance. In Paper I we make an investigation into current generalization bounds with an eye towards practicality and tightness. We find that the current field has some fundamental trade-offs and issues which makes it difficult to find performance guarantees which are both practical to compute and also tight. We conjecture that a novel approach or alternative sets of assumptions might be needed to reach these goals. Therefore, in Paper II we propose and evaluate a new set of assumptions which may be more plausible in real-world settings. We construct a novel setting by taking inspiration from the privileged information framework, introduced in Vapnik & Vashist (2009). Here we assume access to additional information during training and by adding assumptions on this additional information and how it relates to the labels, these latter assumption are made for the input features in most other cases. In addition, we show that our new setting achieves consistent learning as well as empirical results showing performance increases and sample efficiency gains compared to other methods.

Chapter 2

Background

2.1 Unsupervised Domain adaptation

In unsupervised domain adaptation (UDA), we want to learn how to predict outcomes, $Y \in \mathcal{Y}$, from input features, $X \in \mathcal{X}$, and then evaluate that model on some unseen set of data. We have features and labels/target values which have been drawn from an underlying distribution $\mathcal{D}$. This distribution over the product space $\mathcal{X} \times \mathcal{Y}$ is referred to as a *domain*. In UDA, we assume that we have access to $(X, Y) \sim \mathcal{S}$ and $\tilde{X} \sim \mathcal{T}_X$; where $\mathcal{S}$ and $\mathcal{T}$ are called the source and target domains respectively and $\mathcal{T}_X$ is the marginal distribution of features in the target. These quantities will be observed through samples $S = \{x_i, y_i\}_{i=1}^n \sim (\mathcal{S})^n$ and $S'_x = \{\tilde{x}_i\}_{i=1}^m \sim (\mathcal{T}_x)^m$, where $(\mathcal{D})^N$ denotes the distribution of a sample of N datapoints drawn i.i.d. from the domain $\mathcal{D}$. The goal of UDA is to learn a mapping, h, such that we minimize the risk of error when applying h to data from the target domain. More formally, we write the minimization of the target risk as follows:

$$\min_{h \in \mathcal{H}} R_{\mathcal{T}}(h), \quad R_{\mathcal{T}}(h) := \mathbb{E}_{X,Y \sim \mathcal{T}} \left[\ell(h(X), Y) \right], \tag{2.1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function. The loss function is specified so that it measures discrepancy between $h(X)$ and the desired label, Y.

The focus of UDA theory is to find assumptions which enable consistent learning of the mapping, h, or how to upper bound the target risk, $R_{\mathcal{T}}$. (Ben-David et al., 2007; Mansour et al., 2009; Blitzer et al., 2008; Johansson et al., 2019; Wu et al., 2019) As we will see there have been many different approaches tried in this setting.

The UDA problem has been studied for some time, initially it was considered in natural language processing (Hwa, 1999; Chelba & Acero, 2006; Blitzer et al., 2006) as the problem was observed to arise naturally in those settings. An early theoretical treatment of the setting was done in Daumé III & Marcu (2006) based on maximum entropy models. However, the first general treatment is due to Ben-David et al. (2007) where they introduced the idea of a discrepancy metric between the source and target domains. Defining other such metrics

has been a prominent focus in several subsequent works such as e.g. Mansour et al. (2009); Cortes & Mohri (2014) where the authors propose a new such metric and then show the bounds on the target risk implied by using it.

In the next section, we will detail the general form of generalization bounds and how they are produced.

2.2 Guarantees and generalization bounds

In many settings, ensuring good performance of a model is critical to successful deployment. High-stakes settings have this attribute, e.g. autonomous driving and making treatment decision in healthcare settings. If the aim is to guarantee a specific level of model performance we need a bound on the target risk as specified in the previous section. Simply showing acceptable performance on held-out datasets is not a guarantee that the performance will not degrade when applied in other settings. Such a degradation in performance has been observed in e.g. the healthcare setting (Zech et al., 2018), and natural language processing (Jia & Liang, 2017).

As the quantity in (2.1) is written as an expectation over the a priori unknown distribution $\mathcal{T}$ we will have to estimate the risk somehow. The most common and intuitive way to do this is by computing an approximation of this using the sample average which we write as

$$\hat{R}_{\mathcal{T}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\tilde{x}_i), \tilde{y}_i) \tag{2.2}$$

for some sample $\{\tilde{x}_i, \tilde{y}_i\}_{i=1}^{n} \sim (\mathcal{T})^n$. However, as we assume that we do not have access to target labels we have to estimate the target risk with something that we actually can calculate. Therefore, to deal with this complication we use theory to connect the expected target risk to the expected source risk.

Further, we assumed that the distributions of the source and target data were different from each other. Therefore, we need to account for the discrepancy between the two distributions. This can be done in several different ways; to illustrate, we show how this can be done very easily in the following bound:

$$R_{\mathcal{T}}(h) = R_{\mathcal{T}}(h) + R_{\mathcal{S}}(h) - R_{\mathcal{S}}(h) \leq R_{\mathcal{S}}(h) + |R_{\mathcal{T}}(h) - R_{\mathcal{S}}(h)|.$$

The last term, $|R_{\mathcal{T}}(h) - R_{\mathcal{S}}(h)|$, is measuring the discrepancy between the source and target domains, this discrepancy we call the *domain shift*. After this step, we wish to bound the expected source risk with a sample average like in (2.2). However, we still need to account for the error between the expected and empirical risk. This means that a sample generalization term must be added. A common way to estimate the sample generalization error of a classifier is to use statistical learning theory which we will detail in the next section.

Thus, if we have the tools to both relate the source risk to the target risk and connect quantities in expectation to their empirical counterparts, we can express generalization bounds on the following form:

$R_{\mathcal{T}} \leq f(\text{Empirical source risk, Domain shift, Sample generalization error})$.

The specific form of f and the terms it depends on is decided by the theoretical approach taken to the steps detailed above. However, the two main forms are whether the domain shift and sample generalization terms are related through addition or multiplication. We will now go into the main theoretical tools used to solve the two challenges detailed above.

2.2.1 The PAC learning framework

The most prevalent theoretical framework for reasoning about the generalization performance of stochastic models is statistical learning theory, more specifically, a theory called Probably Approximately Correct (PAC) learning (Valiant, 1984). This theory allows us to through assumptions on the model, task and data show that a certain task is learnable, specifically as understood through the PAC lens. What this means is that for a certain task it can be shown to be PAC learnable if we can show that given an algorithm $\mathcal{A}$ and a sample of size n, the algorithm $\mathcal{A}$ returns a model from the model class, $\mathcal{H}$, which has a small average error, ϵ, with high probability, $1-\delta$. This then amounts to that we can show that the risk for a specific model on the given data is smaller than some value $\epsilon > 0$ with confidence level $1-\delta$, where $\delta < 1$, or more formally,

$$\Pr[\mathbb{E}[\ell(h(x), y)] \leq \epsilon] \geq 1 - \delta, \ \forall h \in \mathcal{H}. \tag{2.3}$$

With a formulation on this form we can then use some well known results, often based on concentration inequalities, such as e.g. Hoeffding's inequality to move from an expectation form to an empirical form. This is due to the inequality providing an upper bound on the probability that the loss deviates from its expected value by more than a certain amount. Using these kinds of techniques we can, using application of standard theory (Vapnik, 1998), get bounds like the following. For an i.i.d sample of size m we have that the following holds with probability at least $1-\delta$ for every $h \in \mathcal{H}$:

$$R_{\mathcal{S}}(h) \leq \hat{R}_{\mathcal{S}}(h) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)}. \tag{2.4}$$

The quantity d in the above expression is the so-called Vapnik-Chervonenkis (VC) dimension. The VC dimension is a measure of how complex the family of functions, $\mathcal{H}$, are. An example of a bound on the target risk that is achieved using this framework is the following one from Ben-David et al. (2007)

$$R_{\mathcal{T}}(h) \leq \underbrace{\hat{R}_{\mathcal{S}}(h)}_{\text{Emp. risk}} + \underbrace{\sqrt{\frac{4(d \log \frac{2em}{d} + \log \frac{4}{\delta})}{m}}}_{\text{Sample generalization}} + \underbrace{d_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda}_{\text{Domain shift}}, \tag{2.5}$$

where d is the VC dimension of the $\mathcal{H}$, λ is the sum of the errors on both domains of the best performing classifier $h^* = \arg\min_{h \in \mathcal{H}}(R_{\mathcal{S}}(h) + R_{\mathcal{T}}(h))$, and $d_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \sup_{A \in \{\{x : h(x) = 1\} : h \in \mathcal{H}\}} |\Pr_{\mathcal{S}}[A] - \Pr_{\mathcal{T}}[A]|$ is the $\mathcal{A}$-distance for the characteristic sets of hypotheses in $\mathcal{H}$.

Using this approach we get a bound which holds uniformly over the class of hypotheses $\mathcal{H}$. This is one of the features of PAC learning, the bounds hold for all classifiers in the considered class. However, this can also be a weakness as this produces bounds which must, by definition, hold for the worst classifier imaginable from the class. Depending on the richness of the class this might be arbitrarily limiting. In response to this issue, there is an extension to the PAC framework which does not suffer the same fate which we will detail next.

2.2.2 The PAC-Bayes framework

PAC-Bayes theory is an extension of PAC theory based on using the PAC framework to understand Bayesian classifiers. This way of analysing classifiers was initially proposed by Shawe-Taylor & Williamson (1997), with the first bound being proved by McAllester (1998).

The framework studies generalization of a posterior distribution ρ over hypotheses in $\mathcal{H}$, learned from data, in the context of a prior distribution over hypotheses, π. The generalization error in ρ may be bounded using a divergence between ρ and π as seen in the following classical result due to McAllester (2013).

For a prior π and posterior ρ on $\mathcal{H}$, a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ and any fixed $\gamma, \delta \in (0, 1)$, we have w.p. at least $1 - \delta$ over the draw of m samples from $\mathcal{D}$, with $\mathrm{KL}(p\|q)$ denoting the Kullback-Liebler (KL) divergence between p and q,

$$\mathop{\mathbb{E}}_{h\sim\rho} R_{\mathcal{D}}(h) \leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{h\sim\rho} \hat{R}_{\mathcal{D}}(h) + \frac{\mathrm{KL}(\rho\|\pi) + \ln(\frac{1}{\delta})}{2\gamma(1-\gamma)m} \ . \tag{2.6}$$

As we can see in (2.6), we now have an expression which is stated as an expectation over the posterior distribution ρ. This is in addition to the expectation over the distribution of the data. Thus the bound holds, on average, for classifiers drawn from the posterior ρ. We can interpret ρ as a distribution over the parameters of our trained classifier. The posterior would then be a distribution around the classifier, effectively covering those classifiers which are close in parameter space to our trained classifier. A prominent feature of the framework then, is that we "pay" with another expectation in order to be able to restrict our prediction to a smaller part of the model space, the models which are close to the learned classifier. We consider an additional expectation to be an expense, as it increases computational complexity to a prospective evaluation of the bound.

However, there are some key things to note with this formulation that are advantageous if we want to estimate the quantities in the bound. First, the shape of the prior and posterior distributions are not explicitly stated and can be chosen freely, the bound will still hold. Further, the distribution ρ is something which we learn from the training data. We will next detail another way in which the PAC-Bayes formulation is preferable when the aim is to achieve tighter bounds.

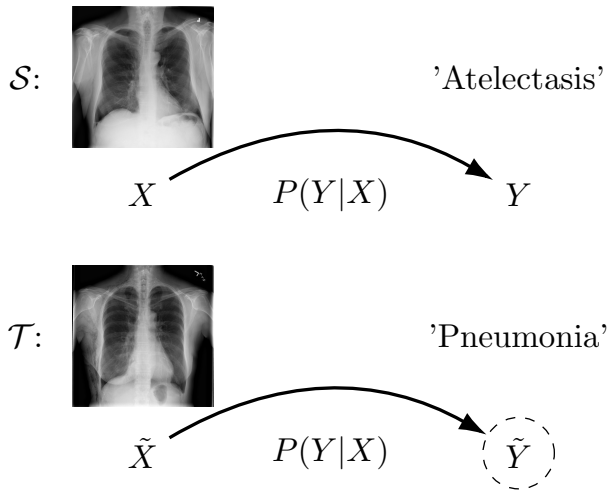

Figure 2.1: A schematic of the general UDA setup. Note that the function from data to label is the same in both the source and target, thereby instantiating the covariate shift assumption. The dashed circle around the target labels $\tilde{Y}$ denotes that they are not an observed quantity.

2.2.3 Data-dependent priors in PAC-Bayes

The KL term in (2.6) grows larger when the prior π and posterior ρ become more dissimilar. This can be the case when the posterior is sensitive to the training data or the prior is badly chosen. To counter this we might try to inform our choice of prior with some of the data we have available. This is called a data-dependent prior and was developed by the work of Ambroladze et al. (2007) and Parrado-Hernández et al. (2012), with an extension to neural networks by Dziugaite et al. (2021).

When we use this type of prior, given that enough data has been used to inform the prior, we will observe a tightening of the resulting bound. This will be due to the KL term being smaller since the prior and posterior are now closer to each other. It is important to note that any data which is used to learn a prior must be independent of the data used to evaluate the bound. If this is not ensured the bound will not hold. However, we should also note that this restriction does not affect which data is used to learn the posterior, ρ.

2.3 Limitations of current UDA theory

In this section, we will talk about the issues in current theoretical work which limits the ability to guarantee consistent learning under realistic assumptions on the UDA problem. We have illustrated the general UDA setup in figure 2.1.

Consistent learning means that we will learn to solve our task in the limit of infinite samples and that we will do so every time. That is, as the sample size n increases the estimates converge in probability to the value that the estimator is designed to estimate. To ensure this, a large swathe of works make assumptions that are quite similar. First, you assume that the underlying function which generates the labels is the same between the domains. This

is called the covariate shift assumption (Shimodaira, 2000), meaning that the data is allowed to change, but not the function labeling the data. We write this as follows:

Assumption 1 (Covariate shift) *For domains $\mathcal{S}, \mathcal{T}$ on $\mathcal{X} \times \mathcal{Y}$, we say that* covariate shift *holds with respect to X if*

$$\exists x : \mathcal{T}(x) \neq \mathcal{S}(x) \;\; \textit{and} \;\; \forall x : \mathcal{T}(Y \mid x) = \mathcal{S}(Y \mid x) \; .$$

This assumption is often made and can hold in many different settings, we often have little reason to believe that the underlying labeling function will change just because the domain has done so.

One might be tempted to think that this assumption is enough, however, this is unfortunately not the case. As shown in Ben-David et al. (2010b) we also need a assumption of coverage of the target domain to have a guarantee of consistent learning. Therefore, we also need the overlapping support assumption to be able to guarantee consistent learning.

Assumption 2 (Domain overlap) *A domain $\mathcal{S}$ overlaps another domain $\mathcal{T}$ with respect to a variable Z on $\mathcal{Z}$ if*

$$\forall z \in \mathcal{Z} : \mathcal{T}(Z = z) > 0 \implies \mathcal{S}(Z = z) > 0 \; .$$

This assumption states that if some datapoint is possible to observe in the target domain we also have a non-zero probability to observe it in the source domain. As should be quite evident, this is quite a strong assumption. We are saying that the target has a probability of already being seen before we apply our model to it. We illustrate this phenomenon in figure 2.2.

To exemplify how assumptions result in limitations on theory we will present some examples from the literature. We start with the following bound due to Cortes et al. (2010)

$$R_{\mathcal{T}} \leq \hat{R}^w_{\mathcal{S}} + 2^{5/4} \sqrt{d_2(\mathcal{T} \| \mathcal{S})} \sqrt[3/8]{\frac{d \log \frac{2ne}{d} + \log \frac{4}{\delta}}{n}}.$$

This bound is an example of an importance weighting bound which bounds the target risk using a weighted empirical source risk, $\hat{R}^w_{\mathcal{S}}$. In this term we re-weight the loss function according to the density ratio $w(x) = \frac{\mathcal{T}(x)}{\mathcal{S}(x)}$ of each sample. For this style of bound we run into issues when the overlap assumption does not hold. Consider the density ratio above; if there is a lack of overlap we may have a data point which only has non-zero density in the target domain. This leads to a division by zero in w and the bound immediately becomes vacuous. The issue is that it is very simple to violate overlap in practice, e.g. learning from black and white images and applying to color images. This inability to handle the non-overlapping case is a weakness we would like to avoid.

So we might come to the conclusion that we should avoid the importance weighting type bounds but still keep the assumptions. This often yields

something akin to the bound we stated in (2.5). We will state a similar bound here due to Ben-David et al. (2010a):

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + 4\sqrt{\frac{2(d\log 2m + \log\frac{2}{\delta})}{m}} + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T}) + \lambda, \tag{2.7}$$

where

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T}) = 2\left(1 - \min_{h,h'\in\mathcal{H}}\left[\frac{1}{m}\sum_{\substack{x\sim(\mathcal{T}_X)^m:\\ h(x)\neq h'(x)}}\mathbb{1}[x] - \frac{1}{m}\sum_{\substack{x\sim(\mathcal{S}_X)^m:\\ h(x)\neq h'(x)}}\mathbb{1}[x]\right]\right).$$

This bound has some qualities that we might take issue with; these are mainly related to the way domain shift is measured. First, the $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ term measures the discrepancy between how much two distinct hypotheses will disagree on the source and target. Intuitively this accounts for the difference between the source and target, but this quantity is not easy to calculate as it requires a minimization over the hypothesis class, $\mathcal{H}$. This is difficult to evaluate as the class can be very large, which is the case for neural network classifiers. This type of of quantity figures in many other works. (Ben-David et al., 2007; Blitzer et al., 2008; Ben-David et al., 2010a; Morvant et al., 2012; Mansour et al., 2009; Redko et al., 2019; Cortes & Mohri, 2014; Cortes et al., 2015).

Second, the λ term, which accounts for the joint optimal error of the best classifier, is not possible to observe with the data available. So if we want a bound that is tight we have to assume that this quantity is small. This non-observable quantity or ones like it is quite common in the literature. (Kuroki et al., 2019; Redko, 2015; Long et al., 2015; Redko et al., 2017; Johansson et al., 2019; Zhang et al., 2019; Dhouib et al., 2020; Shen et al., 2018; Courty et al., 2017; Germain et al., 2013; Dhouib & Redko, 2018; Acuna et al., 2021) As we have seen in this section, the limitations of the current literature are not insubstantial. We therefore see a need to develop theoretical results which do not suffer from these limitations.

2.4 Privileged information

Learning using privileged information (LUPI) is a framework that was first introduced by Vapnik & Vashist (2009). In this setting we use auxiliary information, which we assume we have access to, when training our model. This auxiliary data is in addition to the data and labels available in the regular supervised learning case. In particular, we only have access to this information at the training stage and not when performing inference.

This additional information is called *privileged information* (PI) and can be many different things: residuals, object segmentations, bounding boxes, depth etc. The main goal in this framework is to accelerate the pace of learning. The motivation for why this would be achieved is that in real-world learning we often have students being taught by a teacher. This teacher has better

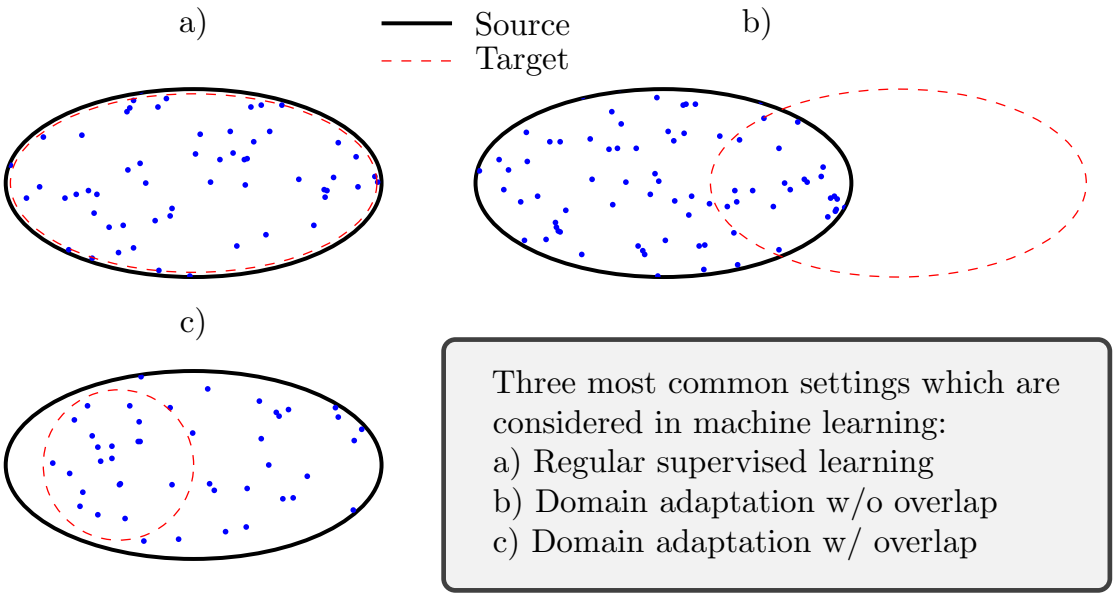

Figure 2.2: Illustration of the different situations regarding common support of data which we might encounter.

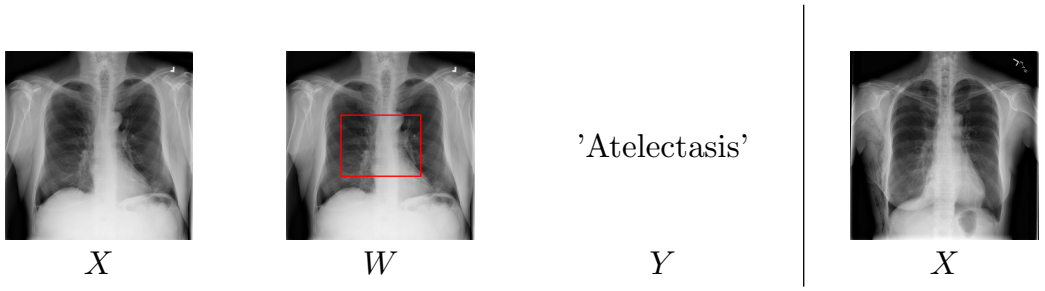

Figure 2.3: An illustration of the PI setting and what data is available. At training time we have access to input features, X, privileged information, W, and labels, Y. At test time we only have the input features, X. Note, that W in this case would be the bounding box.

knowledge about what material, and how it should be presented to the students in order for them to learn the concepts faster. This may be specific explanations, analogies and similar interjections; it is this process that the framework seeks to imitate. So, using this analogy further there would be data (x, y) generated by nature and the privileged information would then be generated by the teacher using the conditional distribution $P(w|x)$ which is assumed to be unknown.

Formally, we assume the existence of some PI, $W \in \mathcal{W}$, which is related to the data through $P(w|x)$. Thus the problem is the following: Given tuples $\{(x_i, w_i, y_i)\}_{i=1}^{N}$ we seek to learn a function f which predicts the outcomes y_i given the data x_i. However, in contrast to regular supervised learning, we only have access to the PI w_i during training. Since we do not assume that we have access to this data at test time, the resulting f cannot explicitly depend on w as an input. Note that we do not assume any specific form or other properties of the PI. An overview of the setting is illustrated in figure 2.3.

The use of PI was initially investigated for use with support vector machines (SVMs), and the framework was later extended to empirical risk minimization (Pechyony & Vapnik, 2010). Methods using PI, which is sometimes called hidden information or side information, has since been applied in many diverse settings such as healthcare (Shaikh et al., 2020), finance (Silva et al., 2010), clustering (Feyereisl & Aickelin, 2012) and image recognition (Vu et al., 2019; Hoffman et al., 2016).

Chapter 3

Summary of Included Papers

We posited in Chapter 1 that we search for ways in which we might achieve successful domain adaptation while *still* having useful guarantees on performance. Both of these goals are attainable in some form for modern neural network models, but rarely together. We have shown in section 2.3 that there are limitations of the current theory that inhibit us from achieving this goal at present. Our first paper deals with illustrating the issue with achieving tractably computable and tight generalization bounds for neural network classifiers. It is quite common that we can find a neural network classifier that performs fairly well on a UDA task, however, there are no realistic guarantees on performance. The second paper proposes a novel set of assumptions, based on privileged information, which we show lead to consistent learning.

3.1 Paper I - Practicality of generalization guarantees for unsupervised domain adaptation with neural networks

In high-stakes scenarios, like the healthcare setting, we would like to have some guarantees on how well our models are going to perform. The most straightforward way of achieving this is through upper bounding the error on the target domain. This can be achieved theoretically in many different ways with varying degrees of usefulness. We can, for instance, trivially state that the error is less than or equal to 1, you cannot be more wrong than all of the time. However, this bound is not so informative so we would like to find better ways of doing this. In Paper I we search the domain adaptation literature for existing bounds which have three properties.

1. Tightness – Is the term a poor approximation? Is it likely to lead to a loose bound?

2. Estimability – Is the term something which we can estimate from observed data?

3. Computability – Can we tractably compute it for real-world data sets and hypothesis classes?

After an extensive literature search we arrive at the conclusion that most available bounds have issues fulfilling these desiderata. Our final selection contains three types of bounds: importance weighting (IW) bounds, bounds containing integral probability metrics (IPM) and PAC-Bayesian bounds. To enable easier comparison we adapt the IW and IPM methods to the PAC-Bayesian framework, thereby creating two novel corollaries to a theorem due to McAllester (2013). These bounds, along with two existing ones due to Germain et al. (2020), are the ones we choose to compute. One of them requires access to target labels to compute and is included for comparison.

We find that without further modification our evaluation results in vacuous bounds due to the sample generalization terms being too large. To remedy this we apply the practice of learning data-dependent priors which entails sacrificing a part of the sample to inform the choice of prior. This tightens the bounds as the sample generalization term, which measures the difference between the prior and posterior distributions, is smaller using this.

We then compute the four different bounds for two image classification tasks, one based on digit classification and one based on X-ray classification. We find that the bound which requires target labels is the tightest, followed by our IW bound which is computable without such information. The other bounds struggle to remain tight even for the simpler digit classification task. We conclude that in cases where our assumptions hold, an importance weighting strategy works well for bounding the error tightly. Further, we conjecture that changing current assumptions will be a way towards a more complete theory explaining out-of-distribution generalization.

Contribution

A. Breitholtz performed the main implementation work, contributed to the writing of the paper and F.D. Johansson supervised the project and contributed to the writing of the paper.

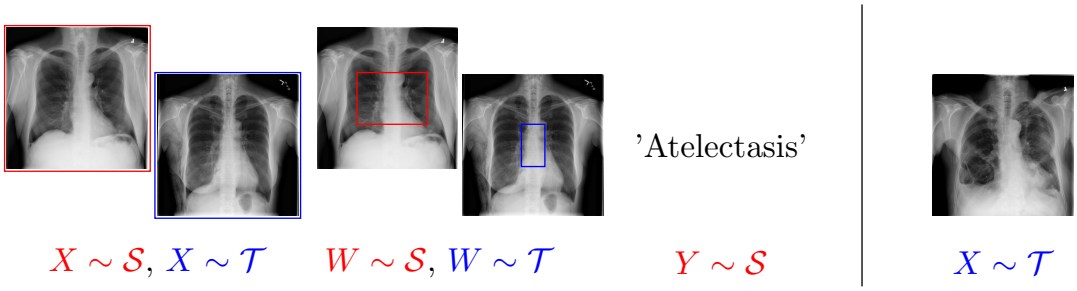

Figure 3.1: An illustration of the data available in the domain adaptation by learning using privileged information (DALUPI) setting. During training, input samples X and privileged information W are available from both source and target domains. Labels Y are only available for inputs from the source domain. At test time, a target sample X is observed.

3.2 Paper II - Unsupervised domain adaptation by learning using privileged information

Taking inspiration from the concept of privileged information (PI), in this paper we propose some changes to the standard set of assumptions. We put forward a version of UDA where we assume access to some privileged information. This privileged information is assumed to be available in both the source and target domains during training time, while at inference we only have access to target features as in regular UDA. We call this setup Domain Adaptation using Learning Using Privileged Information (DALUPI). The general structure of our setting is illustrated in 3.1. We set out to construct theory based on this new structure which ensures consistent learning without the reliance on the overlapping support assumption in the input space. The overlap assumption is often violated in practice and as such it is not ideal to build UDA theory using it.

This new setting enables a very natural way of transferring the model from the source to the target. We simply learn two separate mappings, one from input features to privileged information and one from privileged information to the outcome. This also gives us a simple way to make theory which conforms to this structure, we just learn one mapping after the other. To avoid the overlap assumption in the input space we instead assume overlap w.r.t the PI. Additionally, we assume covariate shift w.r.t the PI similar to what is used in the regular setting. If we add the additional assumption that PI is sufficient for predicting the outcome we show that will have consistent learning. We also propose a bound on the target risk for this setting. We conduct experiments on three different tasks; a synthetic experiment where we investigate how well a model, which is the composition of two separate mappings learned independently, performs when the amount of overlap is varied. The dataset is constructed by inserting a digit in a larger image and having the bounding box around the digit as PI. It shows that a model based on our framework outperforms all other models. We then perform two other experiments, one entity classification task based on the MS-COCO dataset

and pathology classification from chest X-rays. For these experiments we also propose an end-to-end model, based on the Faster R-CNN architecture (Ren et al., 2015). Throughout, the PI we consider are bounding boxes around the region(s) which are informative for the labeling.

From the entity classification task we learn that our method outperforms both the UDA baseline well as performs on par with the model which has been given access to target labels. From the X-ray classification task we learn that the use of PI can yield increased sample efficiency, in line with previous observations. However, the sufficiency of the PI in this task is not obvious, nor guaranteed. We conclude therefore that the DALUPI setting can be beneficial, even when our assumptions are unlikely to hold. In addition, we note that a domain expert will likely be able to judge the sufficiency of PI for tasks like the X-ray classification task we considered.

Contribution

A. Matsson contributed with the implementation of the RCNN model and the COCO and chest X-ray experiments. He also made major contributions to the writing of the paper. A. Breitholtz contributed with the design and implementation of the Synthetic experiment as well as part of the theoretical work. He also made major contributions to the writing of the paper. F.D. Johansson supervised the project and contributed to the theoretical work as well as to the writing of the paper.

Chapter 4

Discussion

In this thesis we have explored approaches to the unsupervised domain adaptation problem; especially as it relates to the connection between what assumptions are made and the resulting guarantees. In Paper I we sought to find tractably computable generalization bounds which are also tight. Through this search we found that the current theory seems to be insufficient to explain model performance without assuming domain overlap. In Paper II, we introduce the notion of using privileged information as a means of achieving provable domain adaptation. Assuming access to this data at training time and that it is sufficient for predicting the label allows us to show that we can achieve consistent learning. In addition, we also find our model performance improves on other methods in cases where our assumptions are likely to hold. Further, we observe increased sample efficiency in our experiments, even in cases where our assumptions are less likely to hold. A limiting factor of Paper II is that it introduces a sufficiency assumption that is not easy to reason about in all cases. The concept of something being sufficient for a prediction would need to be anchored in real-world domain knowledge. However, we argue that the question of sufficiency is, if not easier to answer, more interpretable than other common assumptions. In addition, the experimental evaluation focused on only one form of PI, bounding boxes. The framework proposed is applicable to any form of information and further investigation of other forms of PI would be interesting.

The findings from these papers suggest that there is indeed still a gap in the theory of UDA that needs to be filled. We see introducing new kinds of assumptions as a key way forward for a more rich theory. There are many ways in which this could be instantiated and we will now point out some possible future directions for continuing our work. In most tasks there is a lot of structure inherent to the problem. This originates from the problem formulation, there we define what outcomes we are interested in, what type of data is used and so on. These components often have some sort of overarching way in which their properties relate to human thinking and the way in which we process data. For example, there is often a natural form of hierarchy present in images. Objects like cars, humans and paintings all have component features, the presence of

which are generally needed for identification by a human observer. A human has two eyes, two arms, one torso etc. This part-whole type of hierarchy is inherent to how we understand the composition of objects and often we point to these when classifying objects. A formalization of this type of structure and to what extent we need the parts of an object to be able to classify them could be a way forward.

Another interesting direction for future work is to consider making theory that is more specific than the current literature. While we ideally want a theory to be general enough to capture the behaviour we observe for most situations; one could imagine that our specification is too broad. Perhaps constraining ourselves to reasoning about specific domains or tasks would enable us to find the conditions for consistent generalization within this more specified setting. An example of such a constraint could be to only use images produced by a single type of sensor. This is generally a part of what makes up the domain as it partly influences the generation of the input features. Naturally, our overlap and covariate shift assumptions would likely not hold for arbitrary choices of sensor. However, considering this part as fixed might help further disentangle which parts of the input features are invariant under the remaining domain shift.

Bibliography

David Acuna, Guojun Zhang, Marc T. Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 66–75. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/acuna21a.html`.

Amiran Ambroladze, Emilio Parrado-hernández, and John Shawe-taylor. Tighter pac-bayes bounds. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010a.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility Theorems for Domain Adaptation. In *International Conference on Artificial Intelligence and Statistics*, 2010b.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, pp. 120, Sydney, Australia, 2006. Association for Computational Linguistics. ISBN 978-1-932432-73-2.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning Bounds for Domain Adaptation. *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pp. 8, 2008.

Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2005.05.005. URL `https://www.sciencedirect.com/science/article/pii/S0885230805000276`.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, January 2014. ISSN 0304-3975.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning Bounds for Importance Weighting. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23, pp. 442–450. Curran Associates, Inc., 2010.

Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation Algorithm and Theory Based on Generalized Discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, Sydney NSW Australia, August 2015. ACM. ISBN 978-1-4503-3664-2.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint Distribution Optimal Transportation for Domain Adaptation. *arXiv:1705.08848 [cs, stat]*, October 2017. arXiv: 1705.08848.

Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research (JAIR)*, 26:101–126, 2006. URL `http://hal3.name/docs/#daume06megam`.

Sofien Dhouib, Ievgen Redko, and Carole Lartizien. Margin-aware adversarial domain adaptation with optimal transport. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2514–2524. PMLR, 13–18 Jul 2020.

Sofien Dhouib and Ievgen Redko. Revisiting (ϵ , γ, τ)-similarity learning for domain adaptation. *NeurIPS*, pp. 7408–7417, 2018.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel M. Roy. On the role of data in PAC-Bayes bounds. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, October 2021.

Jan Feyereisl and Uwe Aickelin. Privileged information for data clustering. *Information Sciences*, 194:4–23, 2012.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2016. arXiv: 1505.07818.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. In *International Conference on Machine Learning*, pp. 738–746. PMLR, May 2013. ISSN: 1938-7228.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. PAC-Bayes and Domain Adaptation. *Neurocomputing*, 379:379–397, February 2020. ISSN 09252312. arXiv: 1707.05712.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with Side Information through Modality Hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 826–834, Las Vegas, NV, USA, June 2016. IEEE.

Rebecca Hwa. Supervised grammar induction using training data with limited constituent information. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pp. 73–79, 1999.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL `https://aclanthology.org/D17-1215`.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536. PMLR, 2019.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised Domain Adaptation Based on Source-Guided Discrepancy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4122–4129, July 2019.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *arXiv:1502.02791 [cs]*, May 2015. arXiv: 1502.02791.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. In *Proceedings of the Conference on Learning Theory*, February 2009.

David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pp. 230–234, New York, NY, USA, July 1998. Association for Computing Machinery. ISBN 978-1-58113-057-7.

David A. McAllester. A PAC-Bayesian Tutorial with A Dropout Bound. *arXiv e-prints*, 1307:arXiv:1307.2118, July 2013.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349, November 2012. ISSN 0219-1377, 0219-3116.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. Pac-bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13(1):3507–3531, dec 2012. ISSN 1532-4435.

Dmitry Pechyony and Vladimir Vapnik. On the theory of learnining with privileged information. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Ievgen Redko. *Nonnegative matrix factorization for transfer learning.* PhD thesis, Paris North University, 2015.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical Analysis of Domain Adaptation with Optimal Transport. *arXiv:1610.04420 [cs, stat]*, July 2017. arXiv: 1610.04420.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 849–858. PMLR, 16–18 Apr 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2988–2997. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/saito17a.html`.

Tawseef Ayoub Shaikh, Rashid Ali, and M. M. Sufyan Beg. Transfer learning privileged information fuels CAD diagnosis of breast cancer. *Machine Vision and Applications*, 31(1):9, February 2020.

John Shawe-Taylor and Robert C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, pp. 2–9, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918916.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein Distance Guided Representation Learning for Domain Adaptation. *arXiv:1707.01217 [cs, stat]*, March 2018. arXiv: 1707.01217.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000. ISSN 0378-3758.

Catarina Silva, Armando Vieira, Antonio Gaspar-Cunha, and Joao Carvalho das Neves. Financial distress model prediction using SVM+. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1–7, July 2010.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, pp. 1134–1142, 1984.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, July 2009.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 9 1998. ISBN 0471030031.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7363–7372, 2019.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6872–6881. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/wu19f.html`.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I. Jordan. Bridging Theory and Algorithm for Domain Adaptation. *arXiv:1904.05801 [cs, stat]*, April 2019. arXiv: 1904.05801 version: 1.