Transferable Curriculum for Weakly-Supervised Domain Adaptation

Yang Shu, Zhangjie Cao, Mingsheng Long (✉), and Jianmin Wang
School of Software, Tsinghua University, China
KLiss, MOE; BNRist; Research Center for Big Data, Tsinghua University, China
{shuyang5656,caozhangjie14}@gmail.com {mingsheng,jimwang}@tsinghua.edu.cn

Abstract

Domain adaptation improves a target task by knowledge transfer from a source domain with rich annotations. It is not uncommon that "source-domain engineering" becomes a cumbersome process in domain adaptation: the high-quality source domains highly related to the target domain are hardly available. Thus, weakly-supervised domain adaptation has been introduced to address this difficulty, where we can tolerate the source domain with noises in labels, features, or both. As such, for a particular target task, we simply collect the source domain with coarse labeling or corrupted data. In this paper, we try to address two entangled challenges of weakly-supervised domain adaptation: sample noises of the source domain and distribution shift across domains. To disentangle these challenges, a Transferable Curriculum Learning (TCL) approach is proposed to train the deep networks, guided by a transferable curriculum informing which of the source examples are noiseless and transferable. The approach enhances positive transfer from clean source examples to the target and mitigates negative transfer of noisy source examples. A thorough evaluation shows that our approach significantly outperforms the state-of-the-art on weakly-supervised domain adaptation tasks.

Introduction

Modern deep networks have pushed forward the boundary of various machine perception tasks, at the expenses of large-scale annotated training samples. The high cost of human labeling effectively limits these approaches to many target tasks with insufficient annotations. Thus, there is strong need to leverage or reuse rich labeled data from a different but related source domain. Such a learning paradigm to establish a discriminative model that reduces the underlying distribution shift between domains is known as domain adaptation (Pan and Yang 2010).

Domain adaptation is an important research problem that finds a wide range of application in machine learning (Pan et al. 2011; Duan, Tsang, and Xu 2012; Zhang et al. 2013; Wang and Schneider 2014), computer vision (Saenko et al. 2010; Gong et al. 2012; Hoffman et al. 2014) and natural language processing (Collobert et al. 2011). A rich literature has revealed that deep networks learn distributed representations that disentangle the explanatory factors of variations behind data which can reduce the domain discrepancy (Donahue et al. 2014; Yosinski et al. 2014). In light of this, recent deep domain adaptation approaches embed domain adaptation modules into deep architectures to match feature distributions across domains, yielding evident performance improvement (Tzeng et al. 2014; Long et al. 2015; Ganin and Lempitsky 2015; Tzeng et al. 2015; Long et al. 2016; 2017; Tzeng et al. 2017).

Existing domain adaptation works assume that the source domains are clean datasets with accurate annotations, free of noises. However, this is an ideal scenario. In real domain adaptation problems, we usually have no access to clean and high-quality datasets, which are time consuming and expensive to collect. It is even rarer that such high-quality datasets can be relevant enough to serve as the source domain from which we can leverage useful knowledge to our target task of interest. In contract, we have to collect data from crowdsourcing platform or crawl from Internet or social media. Such datasets are large-scale and to which we have easier access, but are inevitably corrupted with noises.

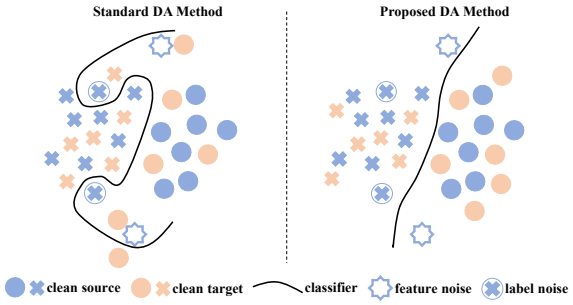


Figure 1: The technical challenge of weakly-supervised domain adaptation. The main idea of the proposed approach is to undo the negative effect of both label and feature noises in the process of distribution matching for domain adaptation.

Thus, it is meaningful to establish methods that can adapt both the representation and classification models from noisy source datasets to our target datasets. This more practical scenario (Figure 1) is known as *weakly-supervised domain adaptation*, which is under-explored compared to standard domain adaptation. In practice, noises mainly present in the two forms of *label noise* and *feature noise*. Label noise refers

to incorrect labels of images, which may be possibly due to errors in manual annotation, low-quality social tagging, label polysemy, or the bias of a crowd-sourcing system etc. Feature noise refers to low-quality pixels of images, which may come from blur, overlap, occlusion, or corruption etc. Thus, weakly-supervised domain adaptation is more general and challenging compared with standard domain adaptation, since the classification model tends to overfit on noisy source data (Zhang et al. 2017), degrading its generalization performance on the target domain. A direct application of existing domain adaptation approaches to the new scenario will simply match the target domain with the entire noisy source domain, resulting in serious negative transfer.

This paper presents a Transferable Curriculum Learning (**TCL**) approach to weakly-supervised domain adaptation, extending from curriculum learning (Bengio et al. 2009) and adversarial learning (Goodfellow et al. 2014). There are two challenges of weakly-supervised domain adaptation: sample noises of the source domain and distribution shift across domains. TCL disentangles these challenges into two alternating optimization subproblems: **1)** Learning with transferable curriculum, which trains our learning model from easy to hard examples and from transferable examples to untransferable examples; **2)** Constructing the transferable curriculum to quantify the transferability of source examples based on their contributions to the target task. As such, the noisy examples are often hard examples and will be selected out and the irrelevant source examples will also be diminished in model training. The final TCL algorithm alternates between these two subproblems to progressively improve both the transferable curriculum and the adaptation model, making the model robust to noises and transferable to target tasks. Experiments show that our method outperforms the state-of-the-art for weakly-supervised domain adaptation problems.

Related Work

Domain adaptation (Pan and Yang 2010) aims to build learning machines that generalize across different but relevant domains. Recent deep domain adaptation methods embed some adaptation modules in deep networks by adding adaptation layers to match the high-order moments of distributions (Tzeng et al. 2014; Long et al. 2015; 2016; 2017), or by exploiting a domain discriminator to distinguish the source and target while learning deep features to confuse the discriminator in an adversarial training paradigm (Ganin and Lempitsky 2015; Tzeng et al. 2015; 2017; Pei et al. 2018; Long et al. 2018). Although these methods have achieved significant improvements, they all assume a clean source domain which is limited and expensive in many real-world applications. State-of-the-art domain adaptation methods may suffer from negative transfer caused by noisy source data in weakly-supervised domain adaptation, which will deteriorate the generalization power of networks trained on the noisy source domain when applied to the target domain.

Learning discriminative models from datasets with noisy labels is an active area of research. Zhang *et al.* (2017) empirically demonstrated that noisy labels will be memorized by DNNs which destroys their generalization capability. One strategy focuses on modeling label noise and class conditional label noise is modeled for binary classification problems (Natarajan et al. 2013). The counterpart for multi-class classification is considered, which introduces an extra noise layer to adapt network outputs to match the noisy label distribution (Sukhbaatar et al. 2014). The noises can be modeled better by learning from privileged information (Vapnik and Izmailov 2015). More recently, a multi-task network is proposed to jointly clean noisy annotations and classify images (Veit et al. 2017). Li *et al.* (2017) proposed a distillation method, utilizing side information from a clean dataset coupled with a knowledge graph. MentorNet is proposed to supervise the training of base networks by learning a data-driven curriculum and assigning appropriate weights to different examples (Jiang et al. 2018). CleanNet is proposed to provide knowledge of label noise with a fraction of manually verified classes (Lee et al. 2018). Another type of methods adjust the loss functions. A bootstrap technique is proposed to alleviate the influence of corrupted labels by augmenting the prediction objective with a notion of consistency (Reed et al. 2014). A dimensionality-driven method is proposed which combines the noisy labels and predicted labels with a local intrinsic dimensionality weight (Ma et al. 2018). Different from these works, we focus on the problem of weakly-supervised domain adaptation, where distribution shift between the source and target domains exists along with noises in the source domain examples.

Weakly-supervised domain adaptation, where the source data constitute noises and distribute dissimilarly to the target data, is an under-explored setting. Probabilistic graphical models are utilized to model the relationships between images, labels and label noises and further integrated into a deep learning system (Xiao et al. 2015). Zeng *et al.* (2014) proposed a scene-specific pedestrian detector by transferring knowledge from a clean dataset. They simply train the model on the target domain of noisy data and thus the clean source domain is needed as an auxiliary dataset. Different from these works, this paper addresses the weakly-supervised domain adaptation problem where the source domain is noisy in labels or features, and the target domain is fully unlabeled. In this setting, the target domain cannot be trained separately due to the lack of supervision. Since the source domain is noisy, it is difficult to enable transfer of only noiseless and relevant source examples.

The proposed Transferable Curriculum Learning (TCL) approach is motivated by curriculum learning (Bengio et al. 2009) which organizes examples in a meaningful order to promote convergence and optimization (Kumar, Packer, and Koller 2010). It prioritizes easier examples of smaller loss by assigning higher weights to them. Different from all previous works, this paper designs a transferable curriculum that simultaneously prioritizes easier *and* transferable examples.

Preliminary on Curriculum Learning

We review the formulation of curriculum learning based on the approach in (Kumar, Packer, and Koller 2010) and (Jiang et al. 2015). Consider a classification task with the training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ of n examples, where $\mathbf{x}_i$ denotes the feature vector of the ith example and $\mathbf{y}_i$ is the associated label over m classes. Denote by $f(\mathbf{x}_i; \boldsymbol{\theta})$ a discriminative

function of a deep network called label classifier, parameterized by $\boldsymbol{\theta}$. Further, let $L(\mathbf{y}_i, f(\mathbf{x}_i; \boldsymbol{\theta}))$ be the loss to quantify the goodness of fit. And a latent weight variable $\mathbf{w} \in \mathbb{R}^n$ is introduced in curriculum learning to optimize the objective:

$$\min_{\boldsymbol{\theta}, \mathbf{w}} E(\boldsymbol{\theta}, \mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} w_i L(\mathbf{y}_i, f(\mathbf{x}_i; \boldsymbol{\theta})) + R(\mathbf{w}; \gamma) \quad (1)$$

where $w_i \in [0, 1]$ is a weight to quantify the contribution for the i-th example. The function R defines a *curriculum*, parameterized by γ. For notation brevity, denote the loss $L(\mathbf{y}_i, f(\mathbf{x}_i; \boldsymbol{\theta})) = \ell_i$.

An alternating minimization algorithm is widely used to solve the above objective (Kumar, Packer, and Koller 2010; Jiang et al. 2015), a procedure where the objective is alternatively minimized over $\boldsymbol{\theta}$ and $\mathbf{w}$, one at a time with the other fixed. When weight $\mathbf{w}$ is fixed, $E(\boldsymbol{\theta}, \mathbf{w})$ is a weighted loss minimized by stochastic gradient descent. When $\boldsymbol{\theta}$ is fixed, we can calculate as $\mathbf{w}^k = \arg\min_{\mathbf{w}} E(\boldsymbol{\theta}^k, \mathbf{w}^{k-1})$ using the most recently updated $\boldsymbol{\theta}^k$ at epoch k. A classic curriculum in (Kumar, Packer, and Koller 2010) is $R(\mathbf{w}; \gamma) = -\gamma \|\mathbf{w}\|_1$, which yields the optimal weight $\mathbf{w}$ as follows,

$$w_i^* = \mathbb{1}(\ell_i \leqslant \gamma), i = 1, \ldots, n, \quad (2)$$

where $\mathbb{1}$ is the indicator function. This update rule interprets the predefined curriculum introduced in (Kumar, Packer, and Koller 2010), known as self-paced learning. First, when updating $\mathbf{w}$ with a fixed model $\boldsymbol{\theta}$, a sample of smaller loss than the threshold γ will be selected as "easy" samples into training ($w_i^* = 1$). Second, when updating $\boldsymbol{\theta}$ with a fixed weight $\mathbf{w}$, the classification model is trained only on the selected "easy" samples. The hyper-parameter $\gamma \geqslant 0$ controls the learning pace and implies the "age" of the model. The model $f(\mathbf{x}; \boldsymbol{\theta})$ can grow up during training, which well imitates the behavior of human learning.

The predefined curriculum specifies a particular sequence of samples with their corresponding weights to be used for self-paced training, where the weights specify the timing and attention to learn each sample. Recent work discovered several curriculums and verified them in many real applications (Jiang et al. 2014; Ma et al. 2017; Fan et al. 2017).

It is worth noting that, when the test data follows similar distribution as the training data, existing curriculum will steadily improve learning. However, when the test data has a distribution shift from the training data, the existing curriculum will be less specified. Due to the distribution shift, the test data will be substantially dissimilar to the training data. That is, even the training samples with smaller loss will be noise-free, they are not necessarily relevant to the learning task of the test data. Thus, there is a need to learn a transferable curriculum able to select samples useful for test data.

Transferable Curriculum Learning

This paper addresses *weakly-supervised domain adaptation*, an under-explored domain adaptation scenario in which the target domain is fully unlabeled and the source domain is partially corrupted with noises in either labels or features. We consider this scenario meaningful and more applicable in practice, since we have much easier access to noisy datasets, e.g., images crawled from social media and search engines are partially annotated with noisy labels and even corrupted with noisy pixels.

The weakly-supervised domain adaptation scenario constitutes a labeled source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ and an unlabeled target domain $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, while the source and target domains follow different distributions $p \neq q$. Note in particular that, we relax the assumption of clean data in standard domain adaptation to that the source domain may be corrupted with noises in either labels or features. Our goal is to train a deep network with a transferable curriculum to eliminate the negative influence of noisy source samples and enable positive transfer of noiseless source samples. The model should also close the domain gap and bound the target risk $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim q}[G_y(G_f(\mathbf{x})) \neq \mathbf{y}]$, by learning transferable features $\mathbf{f} = G_f(\mathbf{x})$ and a robust classifier $\mathbf{y} = G_y(\mathbf{f})$ across noisy source domain and clean target domain.

We may consider tackling the weakly-supervised domain adaptation problem by the curriculum learning introduced in the previous section. In curriculum learning, a curriculum can be specified to select those source samples with smaller losses into training, thus eliminating the negative influence of noisy source samples. However, things become complex when the target domain follows a different distribution from the source domain. First, due to the distribution shift across domains, a classification model trained on the source domain cannot generalize to the target domain. Second, the sample noises introduce difficulty in identifying which fraction of the source samples are transferable to the target task. Thus, the two challenges in distribution shift and sample noises are entangled, making the existing curriculum learning and domain adaptation approaches infeasible.

We present a new transferable curriculum learning (**TCL**) approach to disentangle the challenges behind the sample noises and distribution shift. TCL is an alternating optimization framework comprised of two dependent subproblems: one is learning with a given transferable curriculum and the other is constructing the desirable transferable curriculum.

Learning with Transferable Curriculum

The focus of the first subproblem is learning a domain adaptation model robust to both sample noises and distribution shift, given the transferable curriculum represented by the weighting scheme $w(\mathbf{x}_1^s), \ldots, w(\mathbf{x}_n^s)$. Note that how to construct the curriculum will be described in the next section. Similar to standard curriculum learning (Kumar, Packer, and Koller 2010; Jiang et al. 2015), we will employ this curriculum to train the model from easy samples to hard samples. But unlike standard curriculum learning, we further enforce the model to learn progressively from *transferable* samples to *untransferable* samples. Thus, the transferable curriculum constructed in the second subproblem should tell whether a sample is easy *and* transferable.

Furthermore, the noises in labels and features will introduce a general dataset bias, which cannot be undone without exploiting additional data (Ren et al. 2018). We thus believe that the exploitation of unlabeled target examples by semi-supervised learning is also indispensable. We make use of

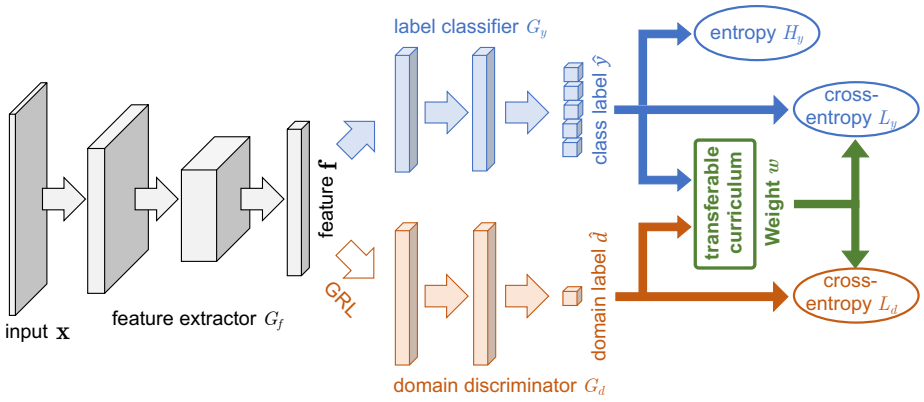


Figure 2: The proposed architecture to learn a transferable curriculum for weakly-supervised domain adaptation. In the diagram, $\mathbf{f}$ is the deep feature extracted by a feature extractor G_f, $\hat{\mathbf{y}}$ is the class label predicted by a label classifier G_y, and $\hat{\mathbf{d}}$ is the domain label predicted by a domain discriminator G_d; The corresponding losses are L_y, L_d and H_y, and GRL is the Gradient Reversal Layer introduced in (Ganin and Lempitsky 2015) for domain adversarial training. In particular, $\mathbf{w}$ is a binary weighting scheme for the transferable curriculum, indicating the selection and timing of each example into training. *Best viewed in color.*

the entropy minimization principle (Grandvalet and Bengio 2005). Letting $\hat{\mathbf{y}}_j^t = G_y(G_f(\mathbf{x}_j^t))$, the entropy loss to quantify the uncertainty of a target example's label predictions is $H_y\left(G_y\left(G_f\left(\mathbf{x}_j^t\right)\right)\right) = -\sum_{c=1}^{m} \hat{y}_{j,c}^t \log \hat{y}_{j,c}^t$. Integrating the entropy loss to the curriculum learning in Eq. (1), we obtain

$$\begin{aligned} E_{G_y} = & \frac{1}{n_s}\sum_{i=1}^{n_s} w(\mathbf{x}_i^s) L_y\left(\mathbf{y}_i^s, G_y\left(G_f\left(\mathbf{x}_i^s\right)\right)\right) \\ & + \frac{1}{n_t}\sum_{j=1}^{n_t} H_y\left(G_y\left(G_f\left(\mathbf{x}_j^t\right)\right)\right), \end{aligned} \tag{3}$$

where L_y is the cross-entropy loss, $w(\mathbf{x}_i^s)$ is the weight in the curriculum that informs whether a source example $\mathbf{x}_i^s$ is noiseless and transferable. Note that given $w(\mathbf{x}_i^s)$, the curriculum regularizer $R\left(\mathbf{w};\gamma\right) = -\gamma\|\mathbf{w}\|_1$ is not involved in training. The model parameters $\boldsymbol{\theta}$ are left out for clarify.

Another technical problem of weakly-supervised domain adaptation is the minimization of distribution shift between the source and target domains. To address this, we use domain adversarial learning (Ganin and Lempitsky 2015) to learn transferable features $\mathbf{f}$ in a two-player minimax game: the first player is a domain discriminator G_d trained to distinguish the feature representations of the source domain from the target domain, and the second player is a feature extractor G_f trained simultaneously to deceive the domain discriminator. Unlike previous work (Ganin and Lempitsky 2015) aligning the entire source domain to the target domain, we only align the target samples with the transferable source samples indicated by the curriculum $w(\mathbf{x}_1^s), \ldots, w(\mathbf{x}_n^s)$. This leads to a novel domain discriminator trained by curriculum learning for weakly-supervised domain adaptation:

$$\begin{aligned} E_{G_d} = & -\frac{1}{n_s}\sum_{i=1}^{n_s} w\left(\mathbf{x}_i^s\right) \log\left(G_d\left(G_f\left(\mathbf{x}_i^s\right)\right)\right) \\ & - \frac{1}{n_t}\sum_{j=1}^{n_t} \log\left(1 - G_d\left(G_f\left(\mathbf{x}_j^t\right)\right)\right). \end{aligned} \tag{4}$$

The curriculum learning enables a progressive growth of the domain discriminator, making it robust to the sample noises from the source domain. In return, the domain discriminator will not align the noisy source samples to the target domain, thus mitigating their negative effects to generalization.

The proposed architecture for learning with transferable curriculum is shown in Figure 2. By weighting the losses of the source classifier G_y and the domain discriminator G_d by the transferable curriculum $w(\mathbf{x}_i^s)$, and combining the entropy minimization criterion, we achieve a new form of domain-adversarial learning. Letting $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_y$, and $\boldsymbol{\theta}_d$ be the parameters of G_f, G_y, and G_d respectively, the objective is trained by a minimax optimization procedure yielding a saddle-point solution $(\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y, \hat{\boldsymbol{\theta}}_d)$:

$$\begin{aligned} (\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y) &= \arg\min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_y} E_{G_y} - E_{G_d}, \\ (\hat{\boldsymbol{\theta}}_d) &= \arg\min_{\boldsymbol{\theta}_d} E_{G_d}. \end{aligned} \tag{5}$$

The simultaneous curriculum learning of label classifier G_y and domain discriminator G_d can effectively disentangle the challenges behind sample noises and distribution shift. Thus, we can improve generalization by harnessing noisy samples and mitigate negative transfer by reducing distribution shift.

Constructing the Transferable Curriculum

Through the aforementioned transferable curriculum learning framework in Eq. (5), the remaining technical problem is how to construct the transferable curriculum, represented by a latent weight $w(\mathbf{x}_i^s)$ for each source example. This constitutes our second subproblem. A straightforward and reasonable solution is to directly adopt the self-paced curriculum in Eq. (2). While such a predefined curriculum can quantify the *easiness* of each source example, it cannot quantify the *transferability* of that example.

The transferability implies the contribution of that source sample to the target task, which can be measured by the sim-

ilarity of that source sample to the target domain. Following this idea, we notice that the domain discriminator G_d trained by curriculum learning in Eq. (5) naturally measures such similarity information. Specifically, the predicted probability $G_d(G_f(\mathbf{x}_i^s))$ indicates the probability of classifying $\mathbf{x}_i^s$ as from the source domain, while $1-G_d(G_f(\mathbf{x}_i^s))$ indicates the probability of classifying $\mathbf{x}_i^s$ as from the target domain. Thus, $1-G_d(G_f(\mathbf{x}_i^s))$ is a good indicator to the transferability (similarity) of a source example $\mathbf{x}_i^s$ to the target domain. To make it comparable to the measure of easiness, which is in terms of the loss values as in Eq. (2), we denote the corresponding cross-entropy loss $\tau_i = -\log\left(1-G_d(G_f(\mathbf{x}_i^s))\right)$ as the measure of transferability.

To achieve a best tradeoff between the easiness and transferability, in this paper, we construct the following linearly-combined weighting scheme as the transferable curriculum:

$$\begin{aligned} w(\mathbf{x}_i^s) &= \mathbb{1}\left(\ell_i + \lambda\tau_i \leqslant \gamma\right) \text{where} \\ \ell_i &= L_y\left(\mathbf{y}_i^s, G_y\left(G_f\left(\mathbf{x}_i^s\right)\right)\right), \\ \tau_i &= -\log\left(1-G_d(G_f(\mathbf{x}_i^s))\right), \end{aligned} \tag{6}$$

where $\lambda > 0$ is a hyper-parameter to tradeoff easiness from transferability. We can verify the validity of the curriculum as follows. First, if a source sample $\mathbf{x}_i^s$ is easy, it will have smaller loss ℓ_i. Second, if that example is transferable (similar) to the target domain, then $G_d(G_f(\mathbf{x}_i^s))$ will approach zero, thus implying smaller loss τ_i. In summary, if a source example is both easy and transferable, it will have a smaller combined loss $\ell_i+\lambda\tau_i$, highly possible to be smaller than the model threshold γ and will be selected into the curriculum learning procedure. Therefore, the latent weighting scheme in Eq. (6) can be served as a valid transferable curriculum.

Alternating Minimax Optimization

Finally, we unify the learning with transferable curriculum and construction of transferable curriculum in an alternating minimax problem, which delivers a saddle-point solution to the proposed transferable curriculum learning (**TCL**) model:

$$\begin{aligned} (\hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\theta}}_y) &= \arg\min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_y} E_{G_y} - E_{G_d}, \\ (\hat{\boldsymbol{\theta}}_d) &= \arg\min_{\boldsymbol{\theta}_d} E_{G_d}, \\ w(\mathbf{x}_i^s) &= \mathbb{1}\left(\ell_i + \lambda\tau_i \leqslant \gamma\right). \end{aligned} \tag{7}$$

TCL can simultaneously filter out the noisy samples from hurting the label classifier and domain discriminator through curriculum learning; and transfer the relevant samples to the target domain through domain-adversarial learning guided by a transferable curriculum. This yields a novel end-to-end deep architecture for weakly-supervised domain adaptation.

Experiments

We evaluate **TCL** with state-of-the-art curriculum schemes and deep domain adaptation methods on three datasets. Code and datasets will be available at `github.com/thuml`.

Setup

Office-31 (Saenko et al. 2010) is a standard dataset for domain adaptation, consisting of 4652 images with 31 classes in 3 distinct domains: *Amazon* (**A**), with images collected from amazon.com, *Webcam* (**W**) and *DSLR* (**D**), with images shot by web camera and digital SLR camera respectively. By permuting the 3 domains, we obtain 6 transfer tasks.

Office-Home (Venkateswara et al. 2017) is a more challenging dataset for visual domain adaptation, consisting of 15,500 images from 65 classes in 4 domains: Artistic (**Ar**), Clip Art (**Cl**), Product (**Pr**) and Real-World (**Rw**). Similarly, we obtain 12 transfer tasks by permuting the 4 domains.

Since these two datasets are almost clean, we create their corrupted counterparts following exactly the protocol in the latest state-of-the-art curriculum learning method MentorNet (Jiang et al. 2018). We create noisy source domains from the original clean dataset in 3 different ways:label corruption, feature corruption, and mixed corruption. For **label corruption**, we change the label of each image uniformly to a random class with probability p_{noise}. For **feature corruption**, each image is corrupted by Gaussian blur and salt-and-pepper noise with probability p_{noise}. As for **mixed corruption**, each image is processed by label corruption and feature corruption with probability $p_{\text{noise}}/2$ independently. In all experiments, we use noisy domains as source domains, and clean domains as target domains. Here p_{noise} is the noise level. All three types of noise can represent the performance of weakly-supervised domain adaptation, while the past literature generally studied the label corruption.

Bing-Caltech (Bergamo and Torresani 2010) was created with Bing and Caltech-256 datasets. The **Bing** dataset was formed by collecting images retrieved by Bing image search for each of the **Caltech-256** category labels. Apart from the statistical differences between Bing images and Caltech images, the Bing dataset consists of rich noises, with presence of multiple objects in the same image, polysemy and caricaturization. We simply use Bing as the noisy source domain and Caltech-256 as the clean target domain. While the experiments on Office-31 and Office-Home are random noisy data, the experiments here represent the performance in real-world weakly-supervised domain adaptation.

We compare Transferable Curriculum Learning (**TCL**) with state-of-the-art deep learning, curriculum learning and domain adaptation methods: **ResNet-50** (He et al. 2016), Self-Paced Learning (**SPL**) (Kumar, Packer, and Koller 2010), **MentorNet** (Jiang et al. 2018), Deep Adaptation Network (**DAN**) (Long et al. 2015), Residual Transfer Network (**RTN**) (Long et al. 2016), Domain Adversarial Neural Network (**DANN**) (Ganin et al. 2016), and Adversarial Discriminative Domain Adaptation (**ADDA**) (Tzeng et al. 2017). We use **ResNet-50** as the network backbone. **SPL** is a classic curriculum learning algorithm that presents training samples in a meaningful order with higher weights assigned to easier samples. **MentorNet** learns data-driven curriculum for training very deep networks from data with corrupted labels. **DAN** learns transferable features by matching high-order moments of cross-domain distributions (Gretton et al. 2012). **RTN** adds the entropy minimization criterion into DAN. **DANN** matches different domains by making them indistinguishable for a domain discriminator. **ADDA** is an asymmetric domain adaptation framework combining discriminative modeling, untied weight sharing and GAN loss.

Table 1: Classification Accuracy (%) on **Office-31** with 40% Corruption of Labels, Features and Both

Method	Label Corruption							Feature Corruption							Mixed Corruption						
	A→W	W→A	A→D	D→A	W→D	D→W	Avg	A→W	W→A	A→D	D→A	W→D	D→W	Avg	A→W	W→A	A→D	D→A	W→D	D→W	Avg
ResNet	47.2	33.0	47.1	31.0	68.0	58.8	47.5	70.2	55.1	73.0	55.0	94.5	87.2	72.5	58.8	39.1	69.3	37.7	75.2	75.5	59.3
SPL	72.6	50.0	75.3	38.9	83.3	64.6	64.1	75.8	59.7	75.7	56.7	93.9	87.8	74.9	77.3	57.5	78.4	47.5	93.4	83.5	72.9
MentorNet	74.4	54.2	75.0	43.2	85.9	70.6	67.2	76.0	60.3	75.5	59.1	93.4	89.9	75.7	76.8	59.5	78.2	52.3	94.4	89.0	75.0
DAN	63.2	39.0	58.0	36.7	71.6	61.6	55.0	73.9	60.2	72.2	59.6	92.5	88.0	74.4	64.4	45.1	71.2	44.7	79.3	78.3	63.8
RTN	64.6	56.2	76.1	49.0	82.7	71.7	66.7	81.0	**64.6**	81.3	62.3	**95.2**	91.0	79.2	76.7	56.9	**84.1**	56.4	93.0	86.7	75.6
DANN	61.2	46.2	57.4	42.4	74.5	62.0	57.3	71.3	54.1	69.0	54.1	84.5	84.6	69.6	69.7	50.0	69.5	49.1	80.1	79.7	66.4
ADDA	61.5	49.2	61.2	45.5	74.7	65.1	59.5	76.8	62.0	79.8	60.1	93.7	89.3	77.0	69.7	54.5	72.4	56.0	87.5	85.5	70.9
TCL	**82.0**	**65.7**	**83.3**	**60.5**	**90.8**	**77.2**	**76.6**	**84.9**	62.3	**83.7**	**64.0**	93.4	**91.3**	**79.9**	**87.4**	**64.6**	83.1	**62.2**	**99.0**	**92.7**	**81.5**

Table 2: Classification Accuracy (%) on **Office-Home** with 40% Mixed Corruption and **Bing-Caltech** with Native Noises

Method	Office-Home													Bing-Caltech
	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg	B→ C
ResNet	27.1	50.7	61.7	41.1	53.8	56.3	40.9	28.0	61.8	51.3	33.0	65.9	47.6	74.4
SPL	32.4	56.0	67.4	41.9	55.3	57.2	47.9	32.9	69.3	60.0	36.2	70.4	52.2	75.3
MentorNet	34.5	57.1	66.7	43.3	56.1	57.6	48.5	34.0	70.2	59.8	37.2	70.4	53.0	75.6
DAN	31.2	52.3	61.2	41.2	53.1	54.6	40.7	30.3	61.5	51.7	36.7	67.4	48.5	75.0
RTN	29.3	57.8	66.3	44.0	**58.6**	58.3	46.0	30.1	67.5	56.3	32.2	69.9	51.4	75.8
DANN	32.9	50.6	60.1	38.6	49.2	50.6	39.9	32.6	60.4	50.5	38.4	67.4	47.6	72.3
ADDA	32.6	52.0	60.6	42.6	53.5	54.3	43.0	31.6	63.1	52.7	37.7	67.5	49.3	74.7
TCL	**38.8**	**62.1**	**69.4**	**46.5**	58.5	**59.8**	**51.3**	**39.9**	**72.3**	**63.4**	**43.5**	**74.0**	**56.6**	**79.0**

We investigate different modules of TCL by the ablation study of its four variants: 1) **TCL-adversarial_w** is the variant by removing the curriculum weight $\{w_i\}_{i=1}^{n_s}$ for the source data on the domain discriminator; 2) **TCL-classifier_w** is the variant by removing the curriculum weight $\{w_i\}_{i=1}^{n_s}$ for the source data on the label classifier; 3) **TCL-easiness** is the variant by removing the easiness term ℓ_i from the curriculum weight in Eq. (6); 4) **TCL-transferability** is the variant by removing the transferability term τ_i from the curriculum weight in Eq. (6).

We follow standard evaluation protocols for unsupervised domain adaptation (Ganin et al. 2016) and use all labeled source examples and unlabeled target examples for training. All deep methods are implemented based on **PyTorch**. We use ResNet-50 pre-trained on the ImageNet dataset (Russakovsky et al. 2015) as our base model, and add a fully-connected bottleneck layer before its classifier layer. Since the dataset is relatively small and the source domain is noisy, we fine-tune only the last residual block of the ResNet-50 model, and train the bottleneck layer, the classifier layer and the domain discriminator from scratch. Before using the curriculum, we pre-train our network on noisy data for a few epochs, which is better than random initialization.

The tradeoff hyper-parameter λ is selected according to magnitudes of the two terms in Eq. (6), and the threshold γ is selected according to the distribution of loss values using cross validation. In each dataset, we simply use the same γ selected from one task for all other tasks under the same noise level. We use mini-batch SGD with momentum of 0.9 and the same learning rate strategy in (Ganin et al. 2016).

Results

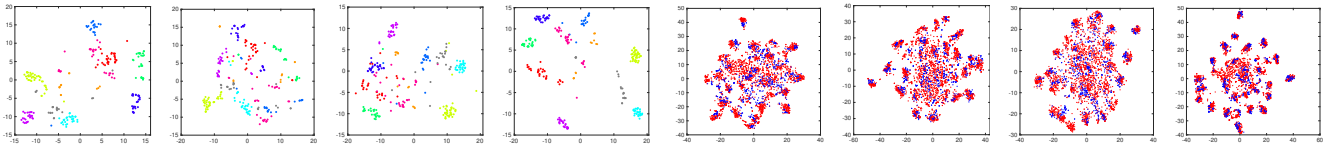
The results on Office-31 under **40%** label corruption, feature corruption, and mixed corruption are shown in Table 1. Further, those of Office-Home under **40%** mixed corruption and Bing-Caltech are reported in Tabel 2. TCL outperforms all the comparison methods on almost all the tasks. In particular, TCL outperforms state-of-the-art deep domain adaptation methods DAN, DANN and ADDA with large margins since these methods suffer from negative transfer and overfitting caused by noisy source examples.

More specifically, ResNet cannot learn a model with high generalization power from the source dataset since it will overfit on the noisy data and be hurt by the distribution shift. RTN performs better than other standard domain adaptation methods, thanks to the entropy minimization criterion that further exploits the clean (unlabeled) target data to harness the noisy (labeled) source data. However, it still suffers from negative transfer and overfitting since all noisy source data participate in training. SPL and MentorNet can learn a robust classifier from the noisy source dataset with a curriculum to eliminate the noisy examples, whilst MentorNet outperforms SPL by learning a data-driven curriculum more robust to noisy examples. However, these methods are still inferior to TCL since they do not have a domain adaptation module to bridge the source and target domains. TCL diminishes the negative impact of noisy source examples on both the label classifier and domain discriminator by learning with a transferable curriculum, which simultaneously mitigate negative transfer caused by noisy source data and promote positive transfer across noiseless source data and clean target data.

Table 3: Accuracy on **Office-31** with 40% Mixed Corruption

Method	Office-31 Mixed Corruption						
	A→W	W→A	A→D	D→A	W→D	D→W	Avg
TCL-adversarial_w	85.9	62.9	82.1	**64.9**	97.4	92.5	81.0
TCL-classifier_w	77.3	63.5	80.5	61.2	96.2	91.5	78.4
TCL-easiness	74.0	63.6	77.3	61.9	96.4	90.4	77.3
TCL-transferability	84.7	63.8	83.1	62.6	97.8	92.2	80.7
TCL	**87.4**	**64.6**	**83.1**	62.2	**99.0**	**92.7**	**81.5**

We perform an investigation across different TCL variants by changing its modules, with results reported in Table 3. TCL outperforms TCL-easiness and TCL-transferability, indicating the reasonable and effective design of the curricu-



(a) DAN (b) DANN (c) RTN (d) TCL (e) DAN (f) DANN (g) RTN (h) TCL

Figure 3: The t-SNE visualization of DAN, DANN, RTN and TCL with class labels (a–d) and domain labels (e–h).

lum weight. TCL outperforms TCL-adversarial_w, validating that the curriculum weights $\{w(\mathbf{x}_i^s)\}_{i=1}^{n_s}$ over the domain adversarial network can diminish the negative effects of noisy examples and mitigate negative transfer. Finally, TCL outperforms TCL-classifier_w with large margin. The reason behind is that the noisy data may severely destroy the source classifier and further deteriorate the curriculum learning procedure, which in turn will decay the quality of the curriculum weights $\{w(\mathbf{x}_i^s)\}_{i=1}^{n_s}$ and break down the progressive training of domain adversarial network. Such a domino effect will cause huge performance crash.

Discussion

Noise Levels: We investigate a wider spectrum of weakly-supervised domain adaptation by varying the level of noises (mixed corruption). Figure 4 shows that TCL outperforms all the comparison methods at each noise level, indicating that TCL can handle noisy source data under various scenarios of weakly-supervised domain adaptation. In particular, we observe that when the noise level is 0%, TCL still performs as well as the state-of-the-art domain adaptation methods DAN, DANN, ADDA and RTN. This proves that TCL can also fit into the standard domain adaptation scenarios.

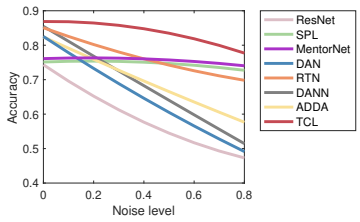


Figure 4: Classification accuracy w.r.t. noise levels.

With increasing levels of noise, the performance of DAN, DANN, and ADDA drops rapidly because noisy source data may severely deteriorate the source classifier and domain adaptation module in existing domain adaptation methods. SPL and MentorNet perform relatively stably since they can down-weigh the negative influence of noisy source data. While this curriculum learning procedure can learn a well-performed classification model from the source noisy dataset, the distribution shift across domains has not been bridged. TCL outperforms all the other methods while yielding high accuracy even at very high noise levels, proving the importance of the transferable curriculum.

Curriculum Quality: We show in Figure 5 the numbers of source samples selected into training (indicated by $w = 1$) when they are label-corrupted, feature-corrupted or clean without corruption. We run TCL on transfer task $\mathbf{A} \rightarrow \mathbf{W}$ under 40% mixed corruption. We observe that the fraction of source data with either label noise or feature noise being selected into training by $w = 1$ are nearly 0%, while the fraction of clean source data being selected into training is approximately 100%. This shows that our transfer curriculum learning mechanism can effectively select out noisy source data and preserve clean data simultaneously.

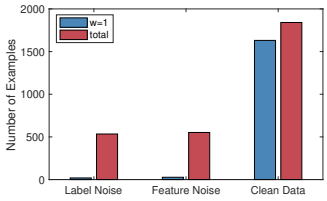


Figure 5: #Examples selected ($w = 1$) for three noise types.

Feature Visualization: We visualize the t-SNE embeddings (Donahue et al. 2014) of the bottleneck representations by DAN, DANN, RTN and TCL on transfer task $\mathbf{A} \rightarrow \mathbf{W}$. Figure 3(a)–3(c) display that the features learned by DAN, DANN, and RTN for different classes are mixed up. Figure 3(e)–3(g) show that the domains are not well aligned while even worse, the target data are aligned to the entire source data with possibly wrong classes, which may cause negative transfer. Figures 3(d) and 3(h) display that TCL can discriminate different classes in both source and target while the target data have similar decision boundary as the source domain. In particular, the noisy data have little influence on knowledge transfer to the target domain. These results validate the efficacy of the transferable curriculum learning.

Conclusion

This paper introduced a new approach to weakly-supervised domain adaptation, an under-explore but more realistic scenario when needing to train from large-scale data with noisy annotations. We proposed a transferable curriculum learning approach to transfer relevant and clean source data while avoiding transfer of noisy or irrelevant source data. The approach yields state-of-the-art results on several real datasets.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2016YFB1000701) and Natural Science Foundation of China (61772299, 61502265, 71690231).

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *ICML*. ACM.

Bergamo, A., and Torresani, L. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *JMLR* 12.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*.

Duan, L.; Tsang, I. W.; and Xu, D. 2012. Domain transfer multiple kernel learning. *TPAMI* 34(3):465–479.

Fan, Y.; He, R.; Liang, J.; and Hu, B. 2017. Self-paced learning: An implicit regularization perspective. In *AAAI*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR* 17(1):2096–2030.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Grandvalet, Y., and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *NIPS*, 529–536. MIT Press.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *JMLR* 13:723–773.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hoffman, J.; Guadarrama, S.; Tzeng, E.; Hu, R.; Donahue, J.; Girshick, R.; Darrell, T.; and Saenko, K. 2014. LSDA: Large scale detection through adaptation. In *NIPS*.

Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014. Self-paced learning with diversity. In *NIPS*.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. 2015. Self-paced curriculum learning. In *AAAI*.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *NIPS*.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*.

Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; and Li, L.-J. 2017. Learning from noisy labels with distillation. In *ICCV*, 1928–1936.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. In *ICML*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *NIPS*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *NIPS*.

Ma, F.; Meng, D.; Xie, Q.; Li, Z.; and Dong, X. 2017. Self-paced co-training. In *ICML*.

Ma, X.; Wang, Y.; Houle, M. E.; Zhou, S.; Erfani, S.; Xia, S.; Wijewickrema, S.; and Bailey, J. 2018. Dimensionality-driven learning with noisy labels. In *ICML*, 3355–3364.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *NIPS*, 1196–1204.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *TKDE*.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *TNNLS*.

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-adversarial domain adaptation. In *AAAI*, 3934–3941.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *ICML*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.

Sukhbaatar, S.; Bruna, J.; Paluri, M.; Bourdev, L.; and Fergus, R. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *CoRR* abs/1412.3474.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.

Vapnik, V., and Izmailov, R. 2015. Learning using privileged information: Similarity control and knowledge transfer. *JMLR* 16:2023–2049.

Veit, A.; Alldrin, N.; Chechik, G.; Krasin, I.; Gupta, A.; and Belongie, S. J. 2017. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, 6575–6583.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *CVPR*.

Wang, X., and Schneider, J. 2014. Flexible transfer learning under support and model shift. In *NIPS*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *NIPS*.

Zeng, X.; Ouyang, W.; Wang, M.; and Wang, X. 2014. Deep learning of scene-specific classifier for pedestrian detection. In *ECCV*, 472–487. Springer.

Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *ICML*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. *ICLR*.