Teaching Semi-Supervised Classifier via Generalized Distillation

Chen Gong[1,2], **Xiaojun Chang**[3], **Meng Fang**[4], **Jian Yang**[1,2]
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] Jiangsu Key Laboratory of Image and Video Understanding for Social Security
[3] Language Technologies Institute, Carnegie Mellon University
[4] Tencent AI Lab
chen.gong@njust.edu.cn

Abstract

Semi-Supervised Learning (SSL) is able to build reliable classifier with very scarce labeled examples by properly utilizing the abundant unlabeled examples. However, existing SSL algorithms often yield unsatisfactory performance due to the lack of supervision information. To address this issue, this paper formulates SSL as a Generalized Distillation (GD) problem, which treats existing SSL algorithm as a learner and introduces a teacher to guide the learner's training process. Specifically, the intelligent teacher holds the privileged knowledge that "explains" the training data but remains unknown to the learner, and the teacher should convey its rich knowledge to the imperfect learner through a specific teaching function. After that, the learner gains knowledge by "imitating" the output of the teaching function under an optimization framework. Therefore, the learner in our algorithm learns from both the teacher and the training data, so its output can be substantially distilled and enhanced. By deriving the Rademacher complexity and error bounds of the proposed algorithm, the usefulness of the introduced teacher is theoretically demonstrated. The superiority of our algorithm to the state-of-the-art methods has also been demonstrated by the experiments on different datasets with various sources of privileged knowledge.

1 Introduction

Semi-Supervised Learning (SSL) [Zhu and Goldberg, 2009] is proposed for solving the real-world problems where the labeled examples are inadequate, yet a large amount of unlabeled examples are relatively easy to obtain. For example, manually annotating web-scale images is intractable due to the unacceptable human labor cost. Acquiring sufficient labeled examples for protein structure categorization is also infeasible as it often takes months of laboratory work for experts to identify a single protein's 3D structure. Fortunately, the massive unlabeled data in these applications are usually available which carry rich distribution information and are helpful for training a reliable classifier.

Regarding how to utilize such abundant unlabeled examples, current SSL algorithms are usually built under *manifold assumption* or *cluster assumption*. Manifold assumption [Belkin *et al.*, 2006; Gong *et al.*, 2015a] assumes that the massive unlabeled examples as well as the scarce labeled examples reveal the underlying manifold of the entire dataset, and the labels of all examples should vary smoothly along this manifold. Cluster assumption [Joachims, 1999; Li and Zhou, 2015] holds that the examples belonging to the same class usually form a cluster in the feature space, and the true decision boundary should fall into the low density region between the well-separated clusters.

From above analyses, we see that existing methods exploit unlabeled examples via different ways to relieve the adverse impact caused by the insufficiency of labeled examples. However, their performances are not stable and are still far from perfect in many situations. Therefore, this paper aims to use machine teaching strategy so that some additional information can be incorporated to aid the training of semi-supervised classifier. Recently, Lopez-Paz *et al.* [Lopez-Paz *et al.*, 2016] proposed a novel Generalized Distillation (GD) framework which integrates the *privileged information* proposed in [Vapnik and Izmailov, 2015] and the *distillation* developed by [Hinton *et al.*, 2015]. In GD, an intelligent "teacher" (i.e. a teaching function) is introduced to transfer its rich knowledge to the imperfect "learner", during which the teacher's knowledge constitutes the privileged information that "explains" the training data to the learner. After that, the learner "receives" the teacher's knowledge by imitating the output of the teacher so that its learning performance can be enhanced, and this process is called "distillation". By this way, a "teacher-learner-collaborative" framework is established which is also very similar to the educational and cognitive process of humans.

The problem studied in this paper is mathematically defined as follows. Suppose we have a set of labeled examples $\mathcal{L} = \{(\mathbf{x}_i, \mathbf{x}_i^*, \mathbf{Y}_i)\}_{i=1}^{l}$ and a set of unlabeled examples $\mathcal{U} = \{(\mathbf{x}_i, \mathbf{x}_i^*, \mathbf{Y}_i)\}_{i=l+1}^{l+u}$ (typically $l \ll u$), where $\mathbf{x}_i$ ($1 \leq i \leq n,\ n = l+u$) are d-dimensional regular data features for the learner, and $\{\mathbf{x}_i^*\}_{i=1}^{n}$ are privileged features that are only

accessible to the teacher. For any $\mathbf{x}_i \in \mathcal{L}$, $\mathbf{Y}_i \in \mathbb{R}^{1\times C}$ (C is the total number of classes) is its label vector with $(\mathbf{Y}_i)_j = Y_{ij}$ being 1 if $\mathbf{x}_i$ belongs to the j-th class, and 0 otherwise. For $\mathbf{x}_i \in \mathcal{U}$, all the elements in the associated $\mathbf{Y}_i$ are initially set to $1/C$ as its groundtruth label is unknown. Given the training set denoted by $\mathcal{Z} = \mathcal{L} \cup \mathcal{U}$, the target of learner is to find a suitable semi-supervised classifier f_s based on $\mathcal{Z} - \{\mathbf{x}_i^*\}_{i=1}^n$ so that the label of an unseen test example $\mathbf{x}_0$ can be perfectly predicted as $y_0 = f_s(\mathbf{x}_0)$. Note that $\{\mathbf{x}_i^*\}_{i=1}^n$ are deducted from $\mathcal{Z}$ as $\{\mathbf{x}_i^*\}_{i=1}^n$ are invisible to the learner. The function of teacher f_t is to generate the soft label $\mathbf{S}_i \in [0,1]^{1\times C}$ for any $\mathbf{x}_i \in \mathcal{U}$ based on the privileged features $\{\mathbf{x}_i^*\}_{i=1}^n$ associated with the training examples. These additional soft labels provide the class prior that is not contained in $\mathcal{Z}$ for the learner, thereby they bias the learned f_s to the output of f_t on $\mathcal{U}$ and also effectively reduce the capacity of f_s's hypothesis space.

The algorithm proposed in this paper is termed "Generalized Distillation for Semi-Supervised Learning" (GDSSL). In GDSSL, we use a popular graph transduction model called Gaussian Field and Harmonic Functions (GFHF) [Zhu *et al.*, 2003] as a teacher, and employ the induction method named Laplacian Regularized Least Squares (LapRLS) [Belkin *et al.*, 2006] as our learner. A transduction method only focuses on labeling the unlabeled examples that are observable in the training set, and does not care about the classification of unseen test data. Therefore, it will probably produce better results on $\mathcal{U}$ than the induction algorithm that pays more attention to the test accuracy than the training accuracy [Gammerman *et al.*, 1998]. As a result, GFHF is adopted as the intelligent teacher f_t to guide the training process of LapRLS (i.e. f_s). Our proposed GDSSL has several merits: Algorithmically, GDSSL enjoys the good property of GD that the privileged information can still be utilized for model training although they are originally withheld from the learner [Lopez-Paz *et al.*, 2016]; Theoretically, we derive the transductive error bound and inductive error bound of the proposed GDSSL, which clearly indicate the usefulness of machine teaching to SSL; and Empirically, by comparing GDSSL with several related state-of-the-art methods, the effectiveness of the proposed method has also been confirmed.

2 Related Work

As this paper aims to harness GD to teach an SSL algorithm, we review the representative prior works on these two topics.

2.1 Semi-Supervised Learning

SSL is specifically designed for the applications where sufficient labeled examples are difficult to obtain while massive unlabeled examples are available. Existing SSL algorithms are usually developed under manifold assumption or cluster assumption.

Cluster assumption assumes that the classes formed by both labeled and unlabeled examples are well-separated, such that the ideal decision boundary falls exactly into the low density area in the feature space. The representative algorithms based on cluster assumption include [Joachims, 1999; Grandvalet and Bengio, 2004; Li *et al.*, 2009; 2016b; Yan *et al.*, 2016]. Manifold assumption based methods explore the geometry of data distribution by postulating that it is supported by a Riemannian manifold. To describe this manifold, a graph is usually built where the vertices correspond to the labeled and unlabeled examples and the edges between the vertices encode the pairwise similarity between them. The label information can then be transmitted from the limited labeled examples to the remaining unlabeled examples through the edges, during which the labels of all examples are required to vary smoothly on the graph. For example, Zhu *et al.* [Zhu *et al.*, 2003], Zhou *et al.* [Zhou *et al.*, 2004] and Gong *et al.* [Gong *et al.*, 2015a] respectively deploy the standard graph Laplacian, normalized graph Laplacian, and deformed graph Laplacian to smooth the labels of examples on the graph. Other works grounded on manifold assumption include [Wang *et al.*, 2009; Belkin *et al.*, 2006; Fang *et al.*, 2014; Gong *et al.*, 2015b; Liu *et al.*, 2017; Gong *et al.*, 2017a; 2017b].

The proposed GDSSL follows manifold assumption, as it is general and includes the cluster assumption as one of its special case [Chapelle *et al.*, 2006]. Apart from harnessing the unlabeled examples for improving the performance like above SSL methods, our GDSSL also incorporates privileged information carried by a teacher via generalized distillation.

2.2 Generalized Distillation

The technique of distillation was initially presented by [Hinton *et al.*, 2015] to reduce the computational cost in test stage for an ensemble of large deep neural networks. Besides, [Vapnik and Izmailov, 2015] proposed a new paradigm named "Learning Using Privileged Information" (LUPI), which incorporates a knowledgable teacher with privileged information and a traditional learner. Here the privileged information, which serve as the additional descriptions to the training data, can only be used by the teacher in training the learner. LUPI has been widely used in many machine learning scenarios such as metric learning [Yang *et al.*, 2016], transfer hashing [Zhou *et al.*, 2016], multi-label learning [You *et al.*, 2017], and deep learning [Hu *et al.*, 2016]. By combining distillation and LUPI into a unified framework, [Lopez-Paz *et al.*, 2016] proposed a novel framework called "Generalized Distillation" (GD). In GD, the teacher transfers its privileged knowledge to the learner, and then the learner "absorbs" these knowledge by mimicking the outputs of the teacher on the training data. Algorithmically, GD has been employed for domain adaptation when the data in source domain are not accessible [Ao *et al.*, 2017]. Practically, Celik *et al.* [Celik *et al.*, 2016] use GD for disease diagnosis when the patients are not willing to reveal their privacy-sensitive data.

To the best of our knowledge, this paper is the first devoted work to adapt GD to solving general SSL problem. Thanks to the efforts of teacher, our method is demonstrated to be more effective than the traditional SSL algorithms from both theoretical and empirical aspects.

3 The Proposed Method

Our GDSSL contains two roles, namely a "teacher" and a "learner". The overall framework of GDSSL is presented in Fig. 1. In training stage, the teacher f_t uses the privileged

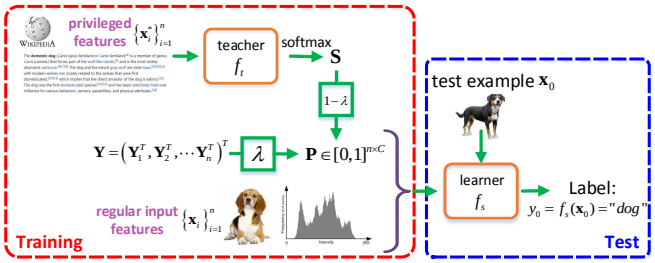

Figure 1: The framework of our algorithm. In training stage, the privileged information (*e.g.* the textual explanations in Wikipedia for the interested objects) is employed by teacher to generate a soft label matrix $\mathbf{S}$, which is further combined with the given labels $\mathbf{Y}$ to form the auxiliary label matrix $\mathbf{P}$. Next, the learner utilizes the regular input features (*e.g.* the color feature of images) as well as $\mathbf{P}$ to train a decision function f_s. In test stage, f_s classifies a test image into the correct class.

information of training data to generate a soft label matrix $\mathbf{S}$ over $\mathcal{Z}$, where each row of $\mathbf{S}$ corresponds to an example. After that, $\mathbf{S}$ and the initial labels in $\mathbf{Y}$ are fused to the auxiliary label matrix $\mathbf{P}$ in a weighted sum way. Then the learner uses $\mathbf{P}$ and the regular input features to establish the decision function f_s. In test stage, a test example $\mathbf{x}_0$ is fed into f_s to get its label $f_s(\mathbf{x}_0)$. Next we detail the training process of GDSSL accomplished by the teacher and the learner.

3.1 Teaching Model

In GDSSL, we use the off-the-shelf transductive method GFHF [Zhu *et al.*, 2003] as the teacher f_t because this method is effective, efficient, and easy to implement. To this end, we build a K-NN graph $\mathcal{G}_t = \langle \mathcal{V}_t, \mathcal{E}_t \rangle$ where $\mathcal{V}_t$ is the vertex set consisted of all n examples in $\mathcal{Z}$, and $\mathcal{E}_t$ is the edge set encoding the similarity between these examples in terms of the privileged features $\{\mathbf{x}_i^*\}_{i=1}^n$. Here the similarity between two examples $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$ is computed by the Gaussian kernel $[W_t]_{ij} = \exp\left(-\left\|\mathbf{x}_i^* - \mathbf{x}_j^*\right\|^2/(2\xi^2)\right)$ with ξ being the kernel width if $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$ are linked by an edge in $\mathcal{G}_t$, and $[W_t]_{ij} = 0$ otherwise. To mathematically quantify $\mathcal{G}_t$, an $n \times n$ adjacency matrix $\mathbf{W}_t$ is computed where its (i,j)-th element is $[W_t]_{ij}$. Therefore, the diagonal degree matrix is defined by $[\mathbf{D}_t]_{ii} = \sum_{j=1}^n [W_t]_{ij}$ and the graph Laplacian matrix is $\mathbf{L}_t = \mathbf{D}_t - \mathbf{W}_t$. By recalling the definition of label vectors $\{\mathbf{Y}_i\}_{i=1}^l$, the objective function of GFHF is [Zhu *et al.*, 2003]

$$\min_{\mathbf{F}} \ \mathrm{tr}(\mathbf{F}^\top \mathbf{L}_t \mathbf{F}) \qquad s.t. \ \mathbf{F}_{\mathcal{L}} = \mathbf{Y}_{\mathcal{L}}, \tag{1}$$

where $\mathbf{Y}_{\mathcal{L}} = \left(\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \cdots, \mathbf{Y}_l^\top\right)^\top$, and $\mathbf{F} = \begin{pmatrix} \mathbf{F}_{\mathcal{L}} \\ \mathbf{F}_{\mathcal{U}} \end{pmatrix} \in \mathbb{R}^{n \times C}$ is the variable to be optimized with each row corresponding to an example in $\mathcal{L}$ or $\mathcal{U}$. By partitioning $\mathbf{L}_t$ as $\mathbf{L}_t = \begin{pmatrix} \mathbf{L}_{\mathcal{L},\mathcal{L}} & \mathbf{L}_{\mathcal{L},\mathcal{U}} \\ \mathbf{L}_{\mathcal{U},\mathcal{L}} & \mathbf{L}_{\mathcal{U},\mathcal{U}} \end{pmatrix}$, the solution of (1) is

$$\mathbf{F}_{\mathcal{L}} = \mathbf{Y}_{\mathcal{L}}, \qquad \mathbf{F}_{\mathcal{U}} = -\mathbf{L}_{\mathcal{U},\mathcal{U}}^{-1} \mathbf{L}_{\mathcal{U},\mathcal{L}} \mathbf{Y}_{\mathcal{L}}. \tag{2}$$

To squash the elements in $\mathbf{F}$ to $[0,1]$, we use a softmax operation to process every $\mathbf{F}_i$ $(i = 1, 2, \cdots, n)$, and obtain

$$\mathbf{S}_{ij} = \frac{\exp([\mathbf{F}_i]_j / T)}{\sum_{i=1}^C \exp([\mathbf{F}_i]_j / T)}, \tag{3}$$

where $[\mathbf{F}_i]_j$ denotes the j-th element in the row vector $\mathbf{F}_i$, and $T > 0$ is the temperature parameter which decides how much we want to soften the $\mathbf{S}$ output by the teacher f_t. Therefore, the element $\mathbf{S}_{ij}$ can be interpreted as the probability of the i-th example belonging to the j-th class.

3.2 Learning Model

This section details how to train an inductive learner $f_s : \mathbb{R}^{d+1} \to \mathbb{R}^C$ with $f_s(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Theta}$ based on $\mathbf{S}$ and $\mathcal{Z} - \{\mathbf{x}_i^*\}_{i=1}^n$. Here $\boldsymbol{\Theta}$ is the coefficient matrix compatible with the augmented $\mathbf{x}$ as $\mathbf{x} := \left(\mathbf{x}^\top \ 1\right)^\top$. According to [Lopez-Paz *et al.*, 2016], the loss function for training the learner f_s on any training example $\mathbf{x}_i \in \mathcal{Z}$ is

$$V(\mathbf{Y}_i, \mathbf{S}_i, f_s(\mathbf{x}_i)) = \lambda \ell(\mathbf{Y}_i, f_s(\mathbf{x}_i)) + (1-\lambda)\ell(\mathbf{S}_i, f_s(\mathbf{x}_i)), \tag{4}$$

where the first term is the traditional *fidelity loss* enforcing the output of $f_s(\mathbf{x}_i)$ to be close to the given label $\mathbf{Y}_i$, and the second term is called *imitation loss* that pushes the learner to generate similar result with the soft label $\mathbf{S}_i$ produced by the teacher. The relative weights of such two terms are governed by a trade-off parameter $\lambda \in [0,1]$. Let $\ell(\mathbf{Y}_i, f_s(\mathbf{x}_i)) = \|\mathbf{Y}_i - f_s(\mathbf{x}_i)\|^2$ be the squared loss in this paper, it can be easily verified that (4) is equivalent to

$$\begin{aligned} V(\mathbf{Y}_i, \mathbf{S}_i, f_s(\mathbf{x}_i)) &= \ell(\lambda \mathbf{Y}_i + (1-\lambda)\mathbf{S}_i, f_s(\mathbf{x}_i)) + Const \\ &\Leftrightarrow \ell(\mathbf{P}_i, f_s(\mathbf{x}_i)), \end{aligned} \tag{5}$$

where $Const$ is irrelevant to f_s, and $\mathbf{P}_i = \lambda \mathbf{Y}_i + (1-\lambda)\mathbf{S}_i$ is the *auxiliary label* in which every element is within $[0,1]$. The cross-entropy loss mentioned in [Lopez-Paz *et al.*, 2016] is not employed here as it makes the learning model do not have closed-form solution. Thereby, inspired by LapRLS [Belkin *et al.*, 2006], the model for pursuing f_s is expressed as

$$\min_{\boldsymbol{\Theta}} \mathcal{J}(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{i=1}^n \ell(\mathbf{P}_i, f_s(\mathbf{x}_i)) + \alpha \mathrm{tr}(\boldsymbol{\Theta}^\top \mathbf{X} \mathbf{L}_s \mathbf{X}^\top \boldsymbol{\Theta}) + \beta \|\boldsymbol{\Theta}\|^2, \tag{6}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ is the training data matrix, and $\mathbf{L}_s$ is the graph Laplacian matrix of learner that can be computed via a similar way with $\mathbf{L}_t$. In (6), the first term is exactly (5), the second term enables the labels of all examples to vary smoothly along the potential manifold, and the third term controls the model complexity to prevent overfitting. Above three terms are weighted by the nonnegative tuning parameters α and β.

To solve (6), we set the derivative of $\mathcal{J}(\boldsymbol{\Theta})$ to $\boldsymbol{\Theta}$ to zero, and then obtain the optimal $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta} = \left(\mathbf{X}\mathbf{X}^\top + \alpha \mathbf{X} \mathbf{L}_s \mathbf{X}^\top + \beta \mathbf{I}\right)^{-1} \mathbf{X} \mathbf{P}, \tag{7}$$

where $\mathbf{I}$ denotes the identity matrix.

Based on the optimized $\boldsymbol{\Theta}$, we may compute the label vector of a test example $\mathbf{x}_0$ as $\mathbf{F}_0 = \mathbf{x}_0^\top \boldsymbol{\Theta}$, and then $\mathbf{x}_0$ is classified into the j-th class with $j = \arg\max_{j' \in \{1, \cdots, C\}} [\mathbf{F}_0]_{j'}$.

4 Theoretical Analyses

For a binary classification problem, the learner's objective function for generating the final decision function becomes

$$\min_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2}\left[\sum_{i=1}^n \left(\boldsymbol{\theta}^\top \mathbf{x}_i - p_i\right)^2 + \alpha \boldsymbol{\theta}^\top \mathbf{X} \mathbf{L}_s \mathbf{X}^\top \boldsymbol{\theta} + \beta \|\boldsymbol{\theta}\|^2\right], \tag{8}$$

where $p_i \in [-1, 1]$ is the target auxiliary label of $\mathbf{x}_i$, and $\boldsymbol{\theta}$ is a d-dimensional coefficient vector to be optimized.

4.1 Semi-Supervised Rademacher Complexity

In computational learning theory, Rademacher complexity is employed to measure the complexity of a class of functions with respect to a certain data distribution. [El-Yaniv and Pechyony, 2009] provide the formulation of Rademacher complexity for SSL, which is

Definition 1. *Let $\mathcal{Z}=\{\mathbf{x}_1,\cdots,\mathbf{x}_l,\mathbf{x}_{l+1},\cdots,\mathbf{x}_{l+u}\}$ be $n=l+u$ examples drawn i.i.d. from a distribution $P_{\mathcal{Z}}$, and $\mathcal{F}$ be a hypothesis space of semi-supervised classifier $f:\mathcal{Z}\to\mathbb{R}$, then the empirical Rademacher complexity of $\mathcal{F}$ is defined by*

$$R_{\mathcal{Z}}(\mathcal{F})=\left(\frac{1}{l}+\frac{1}{u}\right)\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^n\sigma_i f(\mathbf{x}_i)\right)\right], \quad (9)$$

where $\sigma_1,\cdots,\sigma_l$ are independent Rademacher variables randomly chosen from $\{1,-1,0\}$ with the probabilities $\tilde{p}$, $\tilde{p}$, and $1-2\tilde{p}$ ($\tilde{p}\in[0,1/2]$), respectively.

Based on above definition, we have the upper bound of Rademacher complexity of our GDSSL, which is

Theorem 2. *Suppose $\boldsymbol{\theta}^*$ is the optimal solution of* (8)*, and $\mathcal{F}_s$ is the hypothesis space of learner that is guided by a teacher in GDSSL. Then the empirical Rademacher complexity of GDSSL satisfies*

$$R_{\mathcal{Z}}(\mathcal{F}_s)\le\frac{n^2}{\sqrt{\beta}lu}\max_i\|\mathbf{x}_i\|. \quad (10)$$

Proof. Since $\boldsymbol{\theta}^*$ is the minimizer of (8), we have $\mathcal{J}(\boldsymbol{\theta}^*)\le\mathcal{J}(\boldsymbol{\theta}=\mathbf{0})=\frac{1}{2}\sum_{i=1}^n p_i^2\le\frac{n}{2}$. By noticing that every term in (8) is nonnegative, we know $\frac{\beta}{2}\|\boldsymbol{\theta}^*\|^2\le\mathcal{J}(\boldsymbol{\theta}^*)\le\frac{n}{2}$, which leads to $\|\boldsymbol{\theta}^*\|\le\sqrt{\frac{n}{\beta}}$. Besides, from Definition 1 we have

$$\begin{aligned}\frac{lu}{n}R_{\mathcal{Z}}(\mathcal{F}_s)&=\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_s}\sum_{i=1}^n\sigma_i f(\mathbf{x}_i)\right]\\&=\mathbb{E}_\sigma\left[\sup_{f:\|\boldsymbol{\theta}\|\le\sqrt{\frac{n}{\beta}}}\sum_{i=1}^n\sigma_i\langle\boldsymbol{\theta},\mathbf{x}_i\rangle\right]\\&=\mathbb{E}_\sigma\left[\sup_{f:\|\boldsymbol{\theta}\|\le\sqrt{\frac{n}{\beta}}}\left\langle\boldsymbol{\theta},\sum_{i=1}^n\sigma_i\mathbf{x}_i\right\rangle\right]\\&\le\sqrt{\frac{n}{\beta}}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\right\|\right], \quad (11)\end{aligned}$$

where $\mathbb{E}_\sigma(\cdot)$ computes the expectation over $\sigma=(\sigma_1,\cdots,\sigma_n)$, and the last inequality holds due to the Cauchy-Schwarz inequality.

By using Jensen's inequality, we obtain $\mathbb{E}_\sigma[\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\|]=\mathbb{E}_\sigma\left[\left(\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\|^2\right)^{1/2}\right]\le\left(\mathbb{E}_\sigma\left[\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\|^2\right]\right)^{1/2}$. Since $\sigma_1,\cdots,\sigma_n$ are independent, it is straightforward that

$$\begin{aligned}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n\sigma_i\mathbf{x}_i\right\|^2\right]&=\mathbb{E}_\sigma\left[\sum_{i,j}\sigma_i\sigma_j\langle\mathbf{x}_i,\mathbf{x}_j\rangle\right]\\&=\sum_{i\ne j}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\mathbb{E}_\sigma[\sigma_i\sigma_j]+\sum_{i=1}^n\langle\mathbf{x}_i,\mathbf{x}_i\rangle\mathbb{E}_\sigma[\sigma_i^2] \quad (12)\\&=0+2\tilde{p}\sum_{i=1}^n\|\mathbf{x}_i\|^2\le n\max_i\|\mathbf{x}_i\|^2,\end{aligned}$$

which further proves (10) by recalling (11). □

4.2 Error Bounds

For an SSL decision function f, its empirical error on the labeled set $\mathcal{L}$ with size l is $\Re_l(f)=\frac{1}{l}\sum_{i=1}^l\ell(f(\mathbf{x}_i),y_i)$ where y_i is the given label of $\mathbf{x}_i$. The transductive error on the unlabeled set $\mathcal{U}$ with size u is defined by $\Re_u(f)=\frac{1}{u}\sum_{i=l+1}^{l+u}\ell(f(\mathbf{x}_i),y_i)$. Similarly, the training error of f on $\mathcal{L}\cup\mathcal{U}$ is $\Re_n(f)=\frac{1}{n}\sum_{i=1}^n l(f(\mathbf{x}_i),y_i)$. Besides, the inductive error (i.e. generalization error) of f is $\Re(f)=\mathbb{E}_{\mathbf{x}\sim P}[\ell(f(\mathbf{x}),y)]$ where $\mathbf{x}$ is generated from a certain distribution P. Based on above definitions, we introduce two useful lemmas for bounding the transductive error and inductive error of the proposed GDSSL:

Lemma 3. *[El-Yaniv and Pechyony, 2009] Let $\mathcal{F}$ be the hypothesis space of f on the training set $\mathcal{Z}=\mathcal{L}\cup\mathcal{U}$ with $|\mathcal{L}|=l$ and $|\mathcal{U}|=u$. Let $c'=\sqrt{32ln(4e)/3}<5.05$, $Q=1/l+1/u$ and $B=\frac{l+u}{(l+u-1/2)(1-1/(2\max(l,u)))}$. For any fixed γ, with the probability at leat $1-\delta$, we have*

$$\Re_u(f)\le\Re_l^\gamma(f)+\frac{1}{\gamma}R_{\mathcal{Z}}(\mathcal{F})+c'Q\sqrt{\min(l,u)}+\sqrt{\frac{QB}{2}\ln\frac{1}{\delta}}. \quad (13)$$

Lemma 4. *[Bartlett and Mendelson, 2002] Suppose $\mathcal{Z}=\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ are generated i.i.d. from a distribution P, and $\mathcal{F}$ is the hypothesis space, then for any $f\in\mathcal{F}$ with the probability at least $1-\delta$, we have*

$$\Re(f)\le\Re_n(f)+2R_{\mathcal{Z}}(\mathcal{F})+3\sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}. \quad (14)$$

By directly putting the result of (10) into (13) and (14), we can easily obtain the transductive error bound and inductive error bound of our GDSSL, which are

Theorem 5. ***(transductive error bound)*** *Given the notations c', γ, Q, and B defined in Lemma 3 with γ fixed to 1, then with the probability at leat $1-\delta$, the transductive error of the student f_s in GDSSL satisfies*

$$\Re_u(f_s)\le\Re_l(f_s)+\frac{n^2}{\sqrt{\beta}lu}\max_i\|\mathbf{x}_i\|+c'Q\sqrt{l}+\sqrt{\frac{QB}{2}\ln\frac{1}{\delta}}. \quad (15)$$

Theorem 6. ***(inductive error bound)*** *Let $\mathcal{F}_s$ be the hypothesis space of learner f_s that is guided by a teacher, then for any $f_s\in\mathcal{F}_s$, with the probability at least $1-\delta$, we have*

$$\Re(f_s)\le\Re_n(f_s)+\frac{2n^2}{\sqrt{\beta}lu}\max_i\|\mathbf{x}_i\|+3\sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}. \quad (16)$$

From above Theorems 5 and 6, we see that both the transductive error on unlabeled training examples and the inductive error on test examples are upper bounded, so our GDSSL is guaranteed to produce very encouraging performance.

4.3 Usefulness of Teaching

From the general bounds displayed in Eqs. (13) and (14), we see that for a certain training set $\mathcal{Z}$, the Rademacher complexity of hypothesis space $\mathcal{F}$ regarding $\mathcal{Z}$ critically determines the upper bounds for transduction and induction. Next we will theoretically show that the teacher introduced by our GDSSL is able to decrease the Rademacher complexity of the space of the generated decision function, so the transductive error bound and inductive error bound of GDSSL can be effectively reduced when compared with the vanilla LapRLS learner that is not equipped with a teacher.

Suppose the decision function produced by LapRLS is f_{Lap}, which is in the hypothesis space $\mathcal{F}_{Lap}$, then it is straightforward that $|\mathcal{F}_{Lap}| \geq |\mathcal{F}_s|$. This is because the prediction of the learner f_s in GDSSL should be not only consistent with the known labels in $\mathcal{L}$, but also mimic the teacher's label output on $\mathcal{U}$. In contrast, the output of f_{Lap} is only required to be similar to the labels of $\mathcal{L}$. Based on this observation, we have the following theorem:

Theorem 7. *Suppose $\mathcal{F}_s$ and $\mathcal{F}_{Lap}$ are respectively the hypothesis spaces of GDSSL and LapRLS, then given a training set $\mathcal{Z}$, their empirical Rademacher complexities satisfy*

$$R_{\mathcal{Z}}(\mathcal{F}_s) \leq R_{\mathcal{Z}}(\mathcal{F}_{Lap}). \tag{17}$$

Proof. Suppose that $\mathcal{F}_{Lap} = \mathcal{F}_s + \overline{\mathcal{F}}$ where $\overline{\mathcal{F}}$ is the complementary space of $\mathcal{F}_s$ in $\mathcal{F}_{Lap}$, and $Q = \frac{1}{l} + \frac{1}{u}$, then according to (9), we have

$$\begin{aligned}
&R_{\mathcal{Z}}(\mathcal{F}_{Lap}) = QE_\sigma\left[\sup_{f\in\mathcal{F}_{Lap}}\left(\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right)\right] \\
&= QE_\sigma\left[\sup_{f\in\mathcal{F}_s+\overline{\mathcal{F}}}\left(\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right)\right] \\
&\leq QE_\sigma\left[\sup_{f\in\mathcal{F}_s}\left(\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right)\right] + QE_\sigma\left[\sup_{f\in\overline{\mathcal{F}}}\left(\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right)\right] \\
&= R_{\mathcal{Z}}(\mathcal{F}_s) + QE_\sigma\left[\sup_{f\in\overline{\mathcal{F}}}\left(\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right)\right].
\end{aligned} \tag{18}$$

By using the Jensen's inequality, it is obvious that $\mathbb{E}_\sigma\left[\sup_{f\in\overline{\mathcal{F}}}, (\sum_{i=1}^n \sigma_i f(\mathbf{x}_i))\right] \geq \sup_{f\in\overline{\mathcal{F}}} \mathbb{E}_\sigma\left[\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right] = \sup_{f\in\overline{\mathcal{F}}} \sum_{i=1}^n \mathbb{E}(\sigma_i) f(\mathbf{x}_i) = 0$. The last equation holds because $\mathbb{E}(\sigma_i) = 0$ based on the Definition 1. Consequently, the Theorem 7 is proved. □

Therefore, the teacher in our GDSSL plays an important role in reducing the learner's empirical Rademacher complexity, which further leads to the decreased transductive error bound and inductive error bound. As a result, GDSSL consistently performs better than the plain LapRLS algorithm that is the learner of GDSSL.

5 Experimental Results

In this section, we investigate the classification ability of GDSSL on various datasets by using different sources of privileged information. The compared algorithms include:

- State-of-the-art SSL algorithms: Reliable Label Inference via Smoothness Hypothesis (ReLISH) [Gong *et al.*, 2014], Label Prediction via Deformed Graph Laplacian (LPDGL) [Gong *et al.*, 2015a], and Laplacian Regularized Least Squares (LapRLS) [Belkin *et al.*, 2006], among which LapRLS is the basic learner in our GDSSL that is not "taught" by a teacher.
- State-of-the-art algorithms that utilize privileged information: Original SVM+ (SVM+) [Vapnik and Izmailov, 2015], and the improved ℓ_2-SVM+ [Li *et al.*, 2016a].

5.1 Face Recognition: High-Resolution Images as Privileged Information

In this experiment, we investigate the ability of compared methods on face recognition. Specifically, the *ORL* dataset [Cai *et al.*, 2006] is employed, which contains totally 400 face images belonging to 40 subjects. Every image has been respectively resized to the resolutions of 32×32 and 64×64, so the 32×32 face images serve as regular input for the learner while the 64×64 images are treated as privileged information that are used for teaching. We directly re-arrange every two-dimensional image example to a long feature vector in which the elements are grayscale values of the corresponding pixels.

Table 1: Performance of compared methods on *ORL* dataset. The best result is marked in bold. "$\surd$" indicates that GDSSL is significantly better than the corresponding method via paired t-test.

	training accuracy	test accuracy
LapRLS	0.713 ± 0.032 $\surd$	0.640 ± 0.066 $\surd$
ReLISH	0.738 ± 0.021 $\surd$	0.700 ± 0.057 $\surd$
LPDGL	0.754 ± 0.019 $\surd$	0.710 ± 0.065 $\surd$
SVM+	-	0.735 ± 0.066 $\surd$
ℓ_2-SVM+	-	0.738 ± 0.058 $\surd$
GDSSL (ours)	$\mathbf{0.776 \pm 0.016}$	$\mathbf{0.755 \pm 0.049}$

In this paper, every compared method is evaluated by the 5-fold cross validation on each dataset, and the average accuracy over the outputs of the five independent runs are reported to assess the performance of a certain algorithm. Therefore, the training set in *ORL* contains 320 examples, in which we randomly select 80 examples into labeled set $\mathcal{L}$ and the remaining 240 examples constitute the unlabeled set $\mathcal{U}$. The labeled set and unlabeled set in each fold are identical for all the compared methods throughout the experiment.

For fair comparison, we build the same 10-NN graph for the graph-based methods including LapRLS, LPDGL, ReLISH, and our GDSSL. In GDSSL, the parameter λ is set to 0.4 by searching the grid $\{0.2, 0.4, 0.6, 0.8\}$. The temperature parameter T is tuned to 0.01. Besides, the trade-off parameters in (6) are $\alpha = \beta = 0.1$. The regularization parameters C and γ in SVM+ and ℓ_2-SVM+ are set to 1 and 0.01 to achieve the best performance. The results obtained by all the compared methods are presented in Table 1. We see that our GDSSL achieves the highest accuracy among all the compared methods. The training accuracies of SVM+ and ℓ_2-SVM+ are not reported as they are fully supervised methods that are only trained on $\mathcal{L}$, so they cannot be directly compared with other semi-supervised methods that are trained on $\mathcal{Z} = \mathcal{L} \cup \mathcal{U}$. The superiority of GDSSL has also been statistically verified by the paired t-test with the significance level 0.05. Furthermore, by comparing the performances of LapRLS and GDSSL, we see that GDSSL leads the plain learner LapRLS with 6.3% on training accuracy and 11.5% on test accuracy, which firmly demonstrate the usefulness of privileged information and the strength of GD-based teacher.

5.2 Document Categorization: Illustrations as Privileged Information

We apply the proposed GDSSL to text categorization to see its strength in dealing with non-image tasks. To this end, we use a recent *Wikipedia* dataset [Pereira *et al.*, 2014] that contains 2866 documents across 10 categories such as "history", "music", and "biology". Apart from the texts, the articles in this dataset are also accompanied by the illustrative images,

Table 2: Performance of compared methods on *Wikipedia* dataset. The best result is marked in bold. "$\surd$" indicates that GDSSL is significantly better than the corresponding method via paired t-test.

	training accuracy	test accuracy
LapRLS	0.643 ± 0.010 $\surd$	0.613 ± 0.013 $\surd$
ReLISH	0.607 ± 0.032 $\surd$	0.604 ± 0.030 $\surd$
LPDGL	0.634 ± 0.018 $\surd$	0.630 ± 0.020 $\surd$
SVM+	-	0.569 ± 0.031 $\surd$
ℓ_2-SVM+	-	0.597 ± 0.035 $\surd$
GDSSL (ours)	$\mathbf{0.663 \pm 0.006}$	$\mathbf{0.658 \pm 0.018}$

Table 3: Performance of compared methods on *CIFAR100* dataset. The best result is marked in bold. "$\surd$" indicates that GDSSL is significantly better than the corresponding method via paired t-test.

	training accuracy	test accuracy
LapRLS	0.565 ± 0.002 $\surd$	0.438 ± 0.003 $\surd$
ReLISH	0.497 ± 0.003 $\surd$	0.266 ± 0.003 $\surd$
LPDGL	0.558 ± 0.001 $\surd$	0.359 ± 0.005 $\surd$
SVM+	-	0.452 ± 0.001 $\surd$
ℓ_2-SVM+	-	0.414 ± 0.002 $\surd$
GDSSL (ours)	$\mathbf{0.593 \pm 0.002}$	$\mathbf{0.460 \pm 0.002}$

which can be naturally regarded as privileged information to aid the classification of text examples. The image and textual features in each document example are directly provided by [Pereira *et al.*, 2014], namely the text is represented by bag-of-words (BOW) encoding and the associated image is characterized by the SIFT descriptor.

Similar to the above experiment, we also conduct 5-fold cross validation on this *Wikipedia* dataset and see the average outputs of the compared methodologies. The split of training set and test set for each fold has been kept identical for all the algorithms. In the training set of size 2293, we only select 20 examples from each class as labeled and thus the supervision information is very limited for GDSSL to achieve satisfactory performance. At this time, the guidance of teacher would be of great value to make up the shortage of labeled examples.

The accuracies obtained by various algorithms are illustrated in Table 2, which clearly indicate that our GDSSL achieves the best results among all approaches. In this dataset, the SSL methods (*e.g.* LapRLS, ReLISH, LPDGL and GDSSL) perform consistently better than the supervised models such as SVM+ and ℓ_2-SVM+. This is because supervised models are trained only based on the insufficient labeled examples, while SSL methods exploit both labeled and unlabeled examples to build a more powerful classifier.

5.3 Natural Image Classification: Different Feature as Privileged Information

In this section, we evaluate our GDSSL on classifying general images from everyday life. Specifically, a very challenging dataset *CIFAR100* [Krizhevsky and Hinton, 2009] is employed here, which contains 60000 32×32 color images across 100 classes with 600 images per class. Practically, an image can be characterized by more than one feature modalities, and the features from additional modality can be used as privileged information to enhance the classification results. Here we use the features extracted by AlexNet [Krizhevsky *et al.*, 2012] as regular input, and employ the features extracted

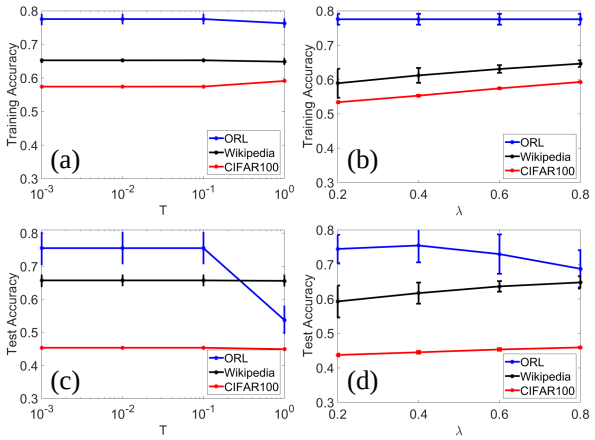

Figure 2: The parametric sensitivity of T and λ. The first column presents the results by changing T, and the second column plots the results w.r.t. the variation of λ.

by VGGNet-16 [Simonyan and Zisserman, 2014] as the privileged information that are only known by the teacher. The 10-NN graph with kernel width $\xi = 10$ is built for LapRLS, ReLISH, LPDGL, and GDSSL.

From the experimental results provided in Table 3, we see that the test accuracies of all compared methods are below 50% because this dataset is very challenging and the labeled examples are too scarce to train a good classifier. However, our GDSSL is still significantly better than other methodologies in terms of both training accuracy and test accuracy.

5.4 Parametric Sensitivity

This section studies how the variations of T in (3) and λ in (4) influence the accuracy of the proposed GDSSL. The training accuracy and test accuracy of GDSSL under different choices of T and λ are observed on the *ORL*, *Wikipedia* and *CIFAR100* datasets. In every dataset, we investigate the effects of T and λ by fixing one of them and then examining the output w.r.t. the change of the other one. From Fig. 2, we see that the training accuracy of GDSSL will not be largely influenced by the choice of these two parameters, however the test accuracy might drop when T gradually increases, so we simply fix T to 0.01 throughout this paper. The determination of λ depends on the "quality" of teacher. If the teacher is able to provide accurate label prior on $\mathcal{U}$, one may choose a small λ to put more emphasize on the output of teacher in (4), otherwise this parameter should be tuned up.

6 Conclusion

To handle the shortage of labeled data in SSL, this paper proposes to incorporate a teacher that brings the additional privileged information to the training process of learner. By designing the imitation loss under the framework of Generalized Distillation [Lopez-Paz *et al.*, 2016], the learner can "study" the knowledge imparted by the teacher as well as from the regular features of training data, and thus it will generate a better classifier than the pure learner without the guidance of a teacher. As the proposed GDSSL is a general framework for teaching a semi-supervised classifier, we plan to apply GDSSL to more SSL algorithms to enhance their classification ability.

Acknowledgments

This research is supported by NSF of China (Nos: 61602246, U1713208 and 61472187), NSF of Jiangsu Province (No: BK20171430), the Fundamental Research Funds for the Central Universities (No: 30918011319), the "Summit of the Six Top Talents" Program (No: DZXX-027), the 973 Program (No: 2014CB349303), and Program for Changjiang Scholars.

References

[Ao *et al.*, 2017] S. Ao, X. Li, and C. Ling. Fast generalized distillation for semi-supervised domain adaptation. In *AAAI*, pages 1719–1725, 2017.

[Bartlett and Mendelson, 2002] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7(11):2399–2434, 2006.

[Cai *et al.*, 2006] D. Cai, X. He, J. Han, and H. Zhang. Orthogonal laplacianfaces for face recognition. *TIP*, 15(11):3608–3614, 2006.

[Celik *et al.*, 2016] Z. Celik, D. Lopez-Paz, and P. McDaniel. Patient-driven privacy control through generalized distillation. *arXiv:1611.08648*, 2016.

[Chapelle *et al.*, 2006] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[El-Yaniv and Pechyony, 2009] R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. *JAIR*, 35(1):193, 2009.

[Fang *et al.*, 2014] Y. Fang, K. Chang, and H. Lauw. Graph-based semi-supervised learning: realizing pointwise smoothness probabilistically. In *ICML*, pages 406–414, 2014.

[Gammerman *et al.*, 1998] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *UAI*, pages 148–155, 1998.

[Gong *et al.*, 2014] C. Gong, D. Tao, K. Fu, and J. Yang. ReLISH: Reliable label inference via smoothness hypothesis. In *AAAI*, pages 1840–1846, 2014.

[Gong *et al.*, 2015a] C. Gong, T. Liu, D. Tao, K. Fu, E. Tu, and J. Yang. Deformed graph Laplacian for semisupervised learning. *TNNLS*, 26(10):2261–2274, Oct. 2015.

[Gong *et al.*, 2015b] C. Gong, D. Tao, K. Fu, and J. Yang. Fick's law assisted propagation for semisupervised learning. *TNNLS*, 26(9):2148–2162, 2015.

[Gong *et al.*, 2017a] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang. Label propagation via teaching-to-learn and learning-to-teach. *TNNLS*, 28(6):1452–1465, 2017.

[Gong *et al.*, 2017b] C. Gong, H. Zhang, J. Yang, and D. Tao. Learning with inadequate and incorrect supervision. In *ICDM*, pages 889–894. IEEE, 2017.

[Grandvalet and Bengio, 2004] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, volume 17, pages 529–536, 2004.

[Hinton *et al.*, 2015] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[Hu *et al.*, 2016] Z. Hu, X. Ma, Z. Liu, E. Hovy, and E. Xing. Harnessing deep neural networks with logic rules. In *ACL*, pages 2410–2420, 2016.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.

[Krizhevsky and Hinton, 2009] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Li and Zhou, 2015] Y. Li and Z. Zhou. Towards making unlabeled data never hurt. *TPAMI*, 37(1):175–188, 2015.

[Li *et al.*, 2009] Y. Li, J. Kwok, and Z. Zhou. Semi-supervised learning using label mean. In *ICML*, pages 633–640. ACM, 2009.

[Li *et al.*, 2016a] W. Li, D. Dai, M. Tan, D. Xu, and L. Gool. Fast algorithms for linear and kernel svm+. In *CVPR*, pages 2258–2266, 2016.

[Li *et al.*, 2016b] Y. Li, J. Kwok, and Z. Zhou. Towards safe semi-supervised learning for multivariate performance measures. In *AAAI*, pages 1816–1822, 2016.

[Liu *et al.*, 2017] Y. Liu, Y. Guo, H. Wang, F. Nie, and H. Huang. Semi-supervised classifications via elastic and robust embedding. In *AAAI*, pages 2294–2300, 2017.

[Lopez-Paz *et al.*, 2016] D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.

[Pereira *et al.*, 2014] J. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36(3):521–535, 2014.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv: 1409.1556v6*, 2014.

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

[Wang *et al.*, 2009] J. Wang, F. Wang, and C. Zhang. Linear neighborhood propagation and its applications. *TPAMI*, 31(9):1600–1615, 2009.

[Yan *et al.*, 2016] Y. Yan, Z. Xu, I. Tsang, G. Long, and Y. Yang. Robust semi-supervised learning through label aggregation. In *AAAI*, pages 2244–2250, 2016.

[Yang *et al.*, 2016] X. Yang, M. Wang, L. Zhang, and D. Tao. Empirical risk minimization for metric learning using privileged information. In *IJCAI*, pages 2266–2272, 2016.

[You *et al.*, 2017] S. You, C. Xu, Y. Wang, C. Xu, and D. Tao. Privileged multi-label learning. *arXiv:1701.07194*, 2017.

[Zhou *et al.*, 2004] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004.

[Zhou *et al.*, 2016] J. Zhou, X. Xu, S. Pan, I. Tsang, Z. Qin, and R. Goh. Transfer hashing with privileged information. *arXiv:1605.04034*, 2016.

[Zhu and Goldberg, 2009] X. Zhu and B. Goldberg. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers, 2009.

[Zhu *et al.*, 2003] X. Zhu, X. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.