Distributionally Robust Learning from Incomplete Data

Amir Najafi,[1,※] Shin-ichi Maeda,[2] Masanori Koyama[2] and Takeru Miyato[2]

[1]Sharif University of Technology, Tehran, Iran [2]Preferred Networks, Inc., Tokyo, Japan

※ Work done while at Preferred Networks, Inc.

Summary

We propose a general framework, SSDRL, that combines distributionally robust learning with semi-supervised learning and includes the following algorithms as its special cases

- Distributionally Robust Learning (DRL) (η=1,only complete data)
- Pseudo-Labeling (PL) (λ→-∞, optimistic estimate of hidden label, ε=0)
- EM algorithm (λ=-1, probabilistic estimate of hidden label by posterior, ε=0)

	DRL (Sinha+,18)	PL (Lee., 13)	VAT (Miyato+, 18)	EM (Dempster+, 77)	SSDRL (Proposed)
Generalization Bound	✓	×	×	×	✓
Convergence Guarantee	✓	×	×	✓	✓
Robustness to Adversaries	✓	×	✓	×	✓
Handling of Unlabeled Data	×	✓	✓	✓	✓

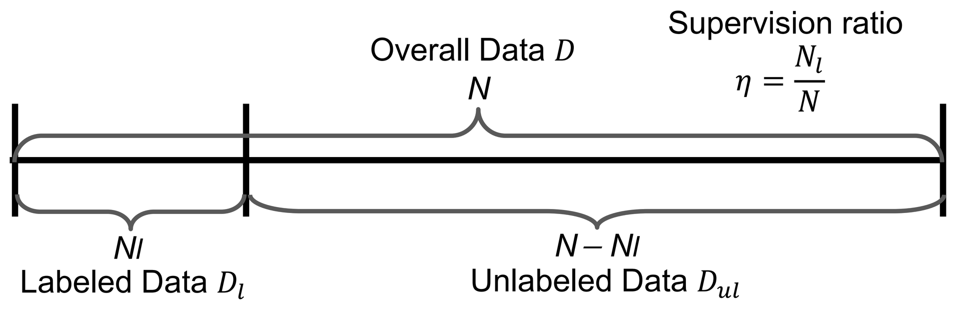
Notations

$p_0(x,y)$: true distribution

$D_l = \{z_1, \cdots, z_{N_l}\}$: Labeled data $z_n = (x_n, y_n) \sim p_0(x,y)$

$D_{ul} = \{x_{N_l+1}, \cdots, x_N\}$: Unlabeled data $x_n \sim p_{0X}(x) \equiv \sum_y p_0(x,y)$

$$\hat{p}(D_l) = \frac{1}{N_l}\sum_{n=1}^{N_l}\delta(z - z_n), \quad \hat{p}(D_{ul}) = \frac{1}{N-N_l}\sum_{n=N_l+1}^{N}\delta(x - x_n)$$: empirical distribution



$B_\epsilon(q) \equiv \{p | W_c(p,q) \le \epsilon\}$: a set of distribution that is close to distribution q where proximity is measured under Wasserstein metric $W_c(p,q)$

$W_c(p,q) \equiv \inf_{\mu \in \prod(p,q)} E_\mu[c(z,z')]$ $c(z,z')$: transportation cost ($c(z,z') \ge 0$, lower semicontinuous, $c(z,z) = 0$) $\prod(p,q)$: a set of joint distributions whose marginal corresponds p and q

Proposed Method: Semi-Supervised Distributionally Robust Learning (SSDRL)

Semi-Supervised Distributionally Robust Learning (SSDRL)

$$\arg\min_\theta \inf_{S \in \hat{P}(D_N)} \left[\sup_{p \in B_\epsilon(S)} E_p[Loss(z;\theta)] + \frac{1-\eta}{\lambda} E_{\hat{p}(D_{ul})}\left[H(S_{y|x})\right] + \gamma\epsilon\right]$$

$$= \arg\min_\theta \left[\left\{\frac{1}{N}\sum_{n=1}^{N_l}\phi_\gamma(z_n;\theta) + \frac{1}{N}\sum_{n=N_l+1}^{N}\mathrm{softmin}_y^\lambda\left(\phi_\gamma\left((x_n,y);\theta\right)\right) + \gamma\epsilon\right\} \equiv R_{SSAR}(\theta;D)\right]$$

$S_{y|x}$: S 's conditional distribution of Y given X

$\hat{P}(D_N) \equiv \{\eta\hat{p}(D_l) + (1-\eta)\hat{p}(D_{ul})Q | Q \in M^X(Y)\}$

$\phi_\gamma(z;\theta) \equiv \sup_{z'}(J(z';z) \equiv Loss(z';\theta) - \gamma c(z',z))$ $(\gamma \ge 0)$

$\mathrm{softmin}_y^\lambda(f_y) \equiv \frac{1}{\lambda}\log\left(\frac{1}{|Y|}\sum_{y \in Y}\exp(\lambda f_y)\right)$

$\eta = 1$ $\eta = 1, \lambda = -1, \epsilon = 0\ (\gamma = +\infty)$ $\lambda = -1, \epsilon = 0\ (\gamma = +\infty)$

DRL

$$\arg\min_\theta \sup_{p \in B_\epsilon(p_0)} E_p[Loss(z;\theta)]$$
$$\approx \arg\min_\theta \sup_{p \in B_\epsilon(\hat{p}(D_l))} E_p[Loss(z;\theta)]$$
$$= \arg\min_\theta \frac{1}{N}\sum_{n=1}^{N_l}\sup_{z'} Loss(z';\theta) - \gamma c(z',z_n)$$

MLE

$$\arg\min_\theta E_{p_0}[Loss(z;\theta)]$$
$$\approx \arg\min_\theta \frac{1}{N_l}\sum_{n=1}^{N_l} Loss(z_n;\theta)$$

EM algorithm

$$\arg\min_\theta \eta E_{p_0}[Loss(z;\theta)] + (1-\eta)\inf_{Q_y \in M^X(Y)}\left(E_{Q_y \times p_{0X}}[Loss(z;\theta) - H(Q_y)]\right)$$
$$\approx \arg\min_\theta \eta\frac{1}{N_l}\sum_{n=1}^{N_l} Loss(z_n;\theta) + (1-\eta)\frac{1}{N-N_l}\sum_{n=N_l+1}^{N}\log\left(\frac{1}{|Y|}\sum_{y \in Y}\exp(-Loss((x_n,y);\theta))\right)$$

Algorithm 1 Stochastic gradient descent for SSDRL

Inputs: $D, \gamma, \lambda, k(\le N), \delta, \alpha, T$

Initialize $\theta_0, t \leftarrow 0$

while $t < T$ **do**

  Randomly select index set I with size k

  for $n \in I$ **do**

    if $n \in I_l \equiv \{1, \cdots, N_l\}$ # Labeled data

      Compute $\hat{z}^*_{n,\theta}$ such that $|\hat{z}^*_{n,\theta} - z^*_{n,\theta}| < \delta$ where $z^*_{n,\theta} = \sup_{z'} Loss(z';\theta) - \gamma c(z', z_n)$

    else # Unlabeled data

      Compute $\hat{z}^*_{n,\theta}(y)$ for each $y \in Y$ such that $|\hat{z}^*_{n,\theta}(y) - z^*_{n,\theta}(y)| < \delta$

      where $z^*_{n,\theta}(y) = \sup_{z'} Loss(z';\theta) - \gamma c(z', (x_n, y))$

    endif

    Compute the gradient $\nabla R^k_{SSAR}(\theta)$

      $\theta_{t+1} \leftarrow \theta_t - \alpha\nabla R^k_{SSAR}(\theta)$

      $t \leftarrow t + 1$

  endfor

endwhile

Output:$\theta^* \leftarrow \theta_T$

F-SSDRL computes adversarial sample only for the *likeliest* y

Computation of adversarial sample for each y

$$\nabla R^k_{SSAR}(\theta) = \frac{1}{k}\sum_{n \in I \cap I_l} g_\theta(\hat{z}^*_{n,\theta}) + \frac{1}{k}\sum_{n \in I \cap I_{ul}}\sum_{y \in Y} q_n(y;\theta) g_\theta(\hat{z}^*_{n,\theta}(y))$$

where $g_\theta(z) \equiv \nabla_\theta Loss(z;\theta)$

$$q_n(y;\theta) = \frac{\exp\left(\lambda J(\hat{z}^*_{n,\theta}(y);(x_n,y))\right)}{\sum_{y' \in Y}\exp\left(\lambda J(\hat{z}^*_{n,\theta}(y');(x_n,y'))\right)}$$

Convergence Guarantee

Assume the loss function is universally differentiable with respect to both parameters z and θ with Lipschitz gradients. Also, assume $\|g_\theta(z)\|_2 \le \sigma$ for some $\sigma \ge 0$ all over $Z \times \Theta$, and $|\lambda| < \infty$. Denote the initial hypothesis as $\theta_0 \in \Theta$, and let $\theta^* \in \Theta$ to be a local minimizer of $R_{SSAR}(\theta;D)$. Also, let $\Delta R \equiv R_{SSAR}(\theta_0;D) - R_{SSAR}(\theta^*;D)$. Then, for a fixed step size α^* as

$$\alpha^* \equiv \frac{1}{\sigma^2}\sqrt{\frac{\Delta R}{T\left(\frac{B}{\sigma^2} + (1-\eta)|\lambda||Y|\right)}},$$

the outputs of Algorithm 1 with parameter set $k = 1, \delta > 0, \alpha = \alpha^*$ after T iterations, satisfy the following inequality:

$$\frac{1}{T}\sum_t^T E\left[\left\|\nabla R^1_{SSAR}(\theta_t)\right\|_2^2\right]$$
$$\le 4\sigma^2\sqrt{\frac{\Delta R}{T}\left(\frac{B}{\sigma^2} + (1-\eta)|\lambda||Y|\right)} + C\delta,$$

where positive constants B and C depend only on γ and the Lipschitz constants associated to $Loss(z;\theta)$.

Generalization Bound

Assume the set of continuous functions $\mathcal{L} \equiv \{Loss(\cdot;\theta) | Loss(\cdot;\theta): Z \to \Re, \|Loss(\cdot;\theta)\|_\infty \le B(\text{for some } B \ge 0), \theta \in \Theta\}$, and $\Phi \equiv \{\phi_\gamma(\cdot;\theta) | \theta \in \Theta\}$. Also assume a partially labeled dataset D which consists of N i.i.d. samples drawn from p_0 where labels can be observed with probability of supervision ratio $\eta \in [0,1]$, independently. For $0 < \delta \le 1$ and $\lambda \le 0$, η satisfies the following condition:

$$\eta \ge MSR_{(\Phi,p_0)}\left(\lambda, 4B\sqrt{\frac{\log\frac{1}{\delta}}{2N}} + 4R^{(\mathrm{SSM})}_{N,(\epsilon,\eta)}(\mathcal{L};p_0)\right)$$

(Newly introduced function $MSR_{(\mathcal{F},P_0)}(\lambda, \text{margin})$ tells us what kind of loss function set $\mathcal{F}$, parameter λ and margin are necessary for the generalization guarantee without observing much labeled data compared with the unlabeled data when the data distribution is P_0. It does not need a restrictive condition like *cluster assumption*.)

Then, with probability at least $1 - \delta$, the following bound holds for all $\epsilon \ge 0$:

$$\sup_{p \in B_\epsilon(p_0)} E_p[Loss(z;\theta^*)] \le \min_{\theta \in \Theta} R_{SSAR}(\theta;D) + 2B\sqrt{\frac{\log\frac{1}{\delta}}{2N}} + 2R^{(\mathrm{SSM})}_{N,(\epsilon,\eta)}(\mathcal{L};p_0)$$

where θ^* is the minimizer of $R_{SSAR}(\theta;D)$.

Definition:

Assume a real-valued function set $\mathcal{F}$ and distribution p_0. Then for $\epsilon \ge 0$ and $\eta \in [0,1]$, Monge Rademacher complexity and SSM Rademacher Complexity of $\mathcal{F}$ according to ϵ-Monge adversaries $\mathrm{A}_\epsilon = \{\forall a: Z \to Z | c(z, a(z)) \le \epsilon, \ \forall z \in Z\}$ are defined as

・Monge Rademacher Complexity

$$R^{(\mathrm{Monge})}_{N,\epsilon}(\mathcal{F};p_0) \equiv \mathrm{E}_{p_0,\sigma}\left[\sup_{f \in \mathcal{F}}\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(\sup_{a \in A_\epsilon} f \circ a(z_n)\right)\right]$$

・Semi-Supervised Monge (SSM) Rademacher Complexity

$$R^{(\mathrm{SSM})}_{N,(\epsilon,\eta)}(\mathcal{F};p_0) \equiv \eta R^{(\mathrm{Monge})}_{N,\epsilon}(\mathcal{F};p_0) + (1-\eta)\sum_{y \in Y} R^{(\mathrm{Monge})}_{N,\epsilon}(\mathcal{F};p_{0X}\delta_y)$$

$z_1, \cdots, z_N \sim p_0$: i.i.d. samples from p_0

$\sigma \in \{-1,+1\}^N$: independent Rademacher random variables

$c(\cdot,\cdot)$: a valid transportation cost

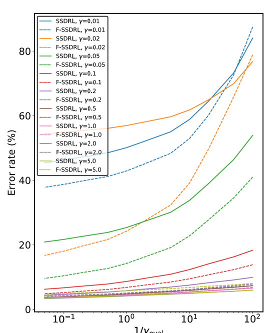
δ_y : the Dirac-delta function over y

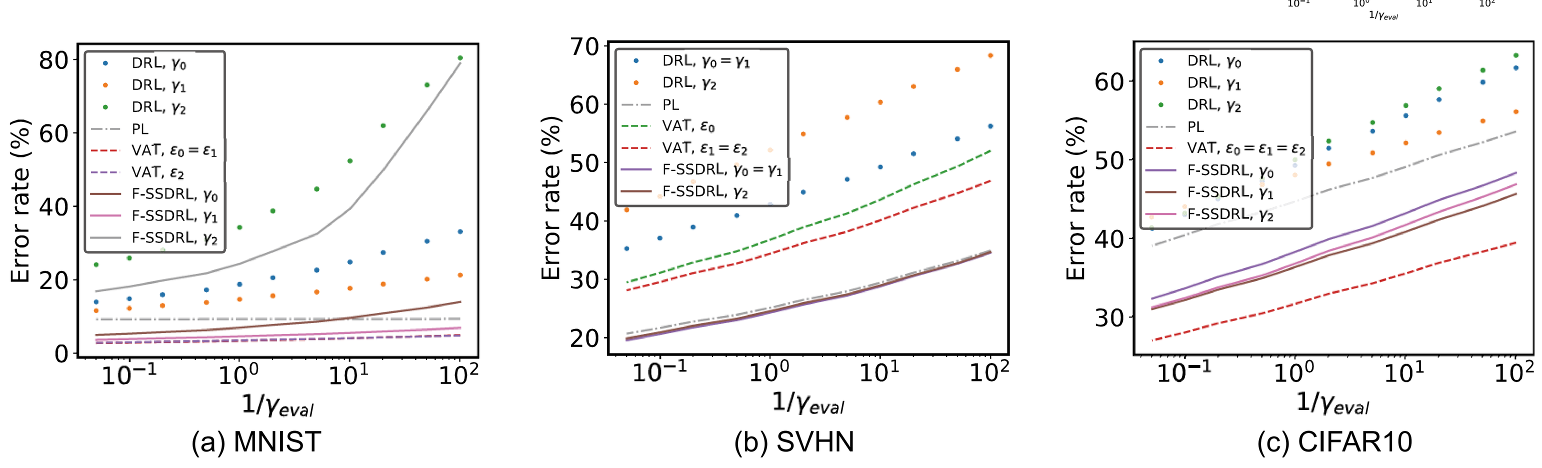
Code is available at https://github.com/pfnet-research/ssdrl

Experiments

We compare the robustness to two kinds of adversarial attacks.

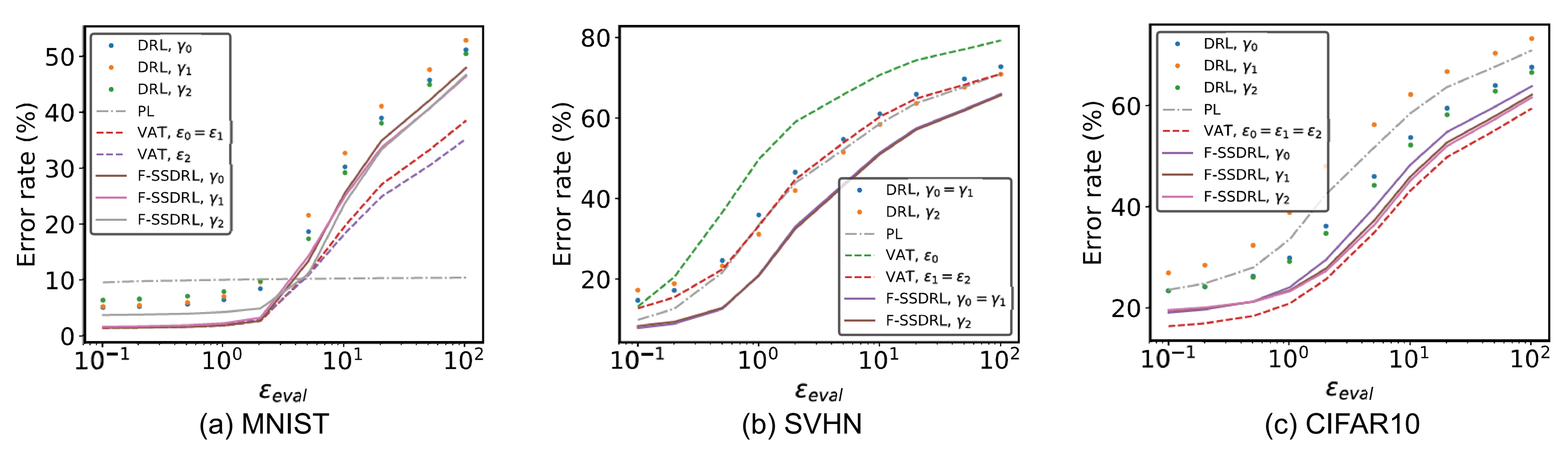
To make a computationally efficient algorithm, we test F-SSDRL that computes adversarial sample only for the *likeliest* y





(a) MNIST (b) SVHN (c) CIFAR10

Figure 1: Robustness to adversarial test examples computed by $\sup_x Loss((x,y^0);\theta) - \gamma\|x - x^0\|_2^2$



(a) MNIST (b) SVHN (c) CIFAR10

Figure 2: Robustness to adversarial test examples computed by projected gradient method, $x^{t+1} = \arg\min_{x \in \{x | \|x - x^0\|_2 \le \epsilon_{eval}\}} \|x - x^t\|_2^2$