Graduation Thesis

Global-Local Regularization Via Distributional Robustness

Phan Viet Hoang, CTTN-CNTT-K63

Computer Science Department
School of Information and Communication Technology.

August 21, 2022

Table of Contents

1. Introduction
2. Proposed approach
3. Experiments
4. Conclusion

Table of Contents

1 Introduction

2 Proposed approach

3 Experiments

4 Conclusion

Adversarial Examples: An Intriguing Phenomenon

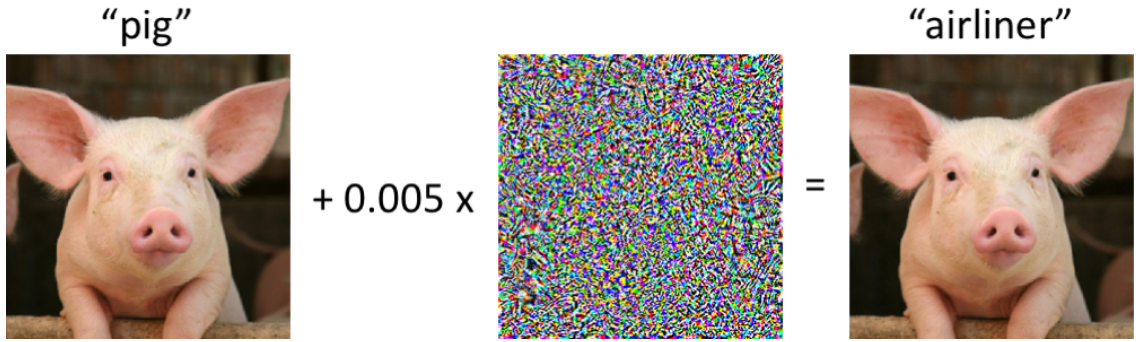

Figure 1: [Goodfellow et al. 2015]

Figure 2: [Athalye et al. 2017]

We want to increase the robustness of machine-learned systems

Wasserstein Distributional Robustness

Wasserstein Distance

Let Ω_S and Ω_T be two separate probability spaces and $\mu \in \mathcal{P}(\Omega_S)$,$\nu \in \mathcal{P}(\Omega_T)$ be two probability measures. For any real number $p \geq 1$, the p -Wasserstein distance between μ and ν that is characterized by the $||\cdot||_p$ norm, is defined in the following way:

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega_S \times \Omega_T} ||\mathbf{x}^s - \mathbf{x}^t||^p d\gamma\left(\mathbf{x}^s, \mathbf{x}^t\right)\right)^{\frac{1}{p}}$$

The Wasserstein distance was proven to be a better way to quantify how distant two distributions are, address the limitations of existing divergence:

- KL (Kullback–Leibler) divergence: $D_{KL}(P\|Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$
- f-divergence: $D_f(P\|Q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x) dx$
- JS (Jensen-Shannon) divergence: $D_{JS}^{\alpha,\beta}(P,Q) = \alpha D_{KL}(P\|\alpha P + \beta Q) + \beta D_{KL}(Q\|\alpha P + \beta Q)$

Wasserstein Distributional Robustness

Distributional robustness (primal form)

Consider a generic Polish space S endowed with a distribution $\mathbb{P}$. Let $r : S \to \mathbb{R}$ be a real-valued (risk) function and $c : S \times S \to \mathbb{R}_+$ be a cost function. Distributional robustness setting aims to find the distribution $\tilde{\mathbb{P}}$ in the vicinity of $\mathbb{P}$ and maximizes the risk in the expectation form:

$$\sup_{\tilde{\mathbb{P}}:\mathcal{W}_c\left(\mathbb{P},\tilde{\mathbb{P}}\right)<\epsilon} \mathbb{E}_{\tilde{Z}\sim\tilde{\mathbb{P}}}\left[r\left(\tilde{Z}\right)\right], \tag{1}$$

where $\epsilon > 0$ and $\mathcal{W}_c\left(\mathbb{P},\tilde{\mathbb{P}}\right) := \inf_{\gamma\in\Gamma\left(\mathbb{P},\tilde{\mathbb{P}}\right)} \int c d\gamma$ denotes an optimal transport (OT) or a Wasserstein (WS) distance with the set of couplings $\Gamma\left(\mathbb{P},\tilde{\mathbb{P}}\right)$ whose marginals are $\mathbb{P}$ and $\tilde{\mathbb{P}}$.

Wasserstein Distributional Robustness

Distributional robustness (dual form)

Assume that $r \in L^1(\mathbb{P})$ is upper semi-continuous and the cost c is a non-negative and continuous function satisfying $c(Z, \tilde{Z}) = 0$ iff $Z = \tilde{Z}$, previous work[1,2] showed the *dual* form for Equation. (1) is:

$$\inf_{\lambda \geq 0} \left\{ \mathbb{E}_{Z \sim \mathbb{P}} [\sup_{\tilde{z}} \left\{ r\left(\tilde{Z}\right) - \lambda c\left(\tilde{Z}, Z\right) \right\}] \right\} \tag{2}$$

[1] Jose Blanchet and Karthyek Murthy. "Quantifying distributional model risk via optimal transport". In: *Mathematics of Operations Research* 44.2 (2019), pp. 565–600.

[2] Aman Sinha, Hongseok Namkoong, and John Duchi. "Certifying Some Distributional Robustness with Principled Adversarial Training". In: *International Conference on Learning Representations*. 2018.

Limitations

The fact that r engages only $\tilde{Z} = \left(\tilde{X}, \tilde{Y}\right) \sim \tilde{\mathbb{P}}$ certainly restricts the modeling capacity of (2). The reasons are as follows:

- Firstly, for each anchor Z, the most challenging sample $\tilde{Z}$ is currently defined as the one maximizing $\sup_{\tilde{Z}} \left(r(\tilde{Z}) - \lambda c\left(Z, \tilde{Z}\right)\right)$, where $r(\tilde{Z})$ is inherited from the primal form (1). Hence, it is not suitable to express the risk function r engaging both Z and $\tilde{Z}$ (e.g., Kullback-Leibler divergence $KL\left(p\left(\tilde{Z}\right) \| p(Z)\right)$ between the predictions for Z and $\tilde{Z}$ as in TRADES[3])
- Secondly, it is also *impossible* to inject a *global regularization term* involving a batch of samples $\tilde{Z}$ and Z.

[3] Hongyang Zhang et al. "Theoretically principled trade-off between robustness and accuracy". In: *Proceedings of ICML*. PMLR. 2019, pp. 7472–7482.

Contributions

Overall, our contributions can be summarized as follows:

1. Propose a rich OT based DR framework, named **G***lobal*-**L***ocal* **O***ptimal* **T***ransport based* **D***istributional* **R***obustness* (GLOT-DR), which enrichs the general framework by enforcing both local and global regularization terms.
2. Propose a closed-form solution for our GLOT-DR without involving the dual form (i.e., equation (2)). Here we note that the dual form (2) is *not computationally convenient* to solve due to the minimization over λ.
3. Conduct comprehensive experiments to compare our GLOT-DR to state-of-the-art baselines in DG, DA, SSL, and AML which empirically proves that both of the introduced local and global regularization terms advance existing methods across various scenarios.

Table of Contents

1 Introduction

2 Proposed approach

3 Experiments

4 Conclusion

Our Framework

Assume that we have *multiple labeled source domains* with the *data/label* distributions $\left\{\mathbb{P}_k^S\right\}_{k=1}^K$ and a *single unlabeled target domain* with the *data* distribution $\mathbb{P}^T$.

- For the k-th source domain, we draw a batch of B_k^S examples as $\left(X_{ki}^S, Y_{ki}^S\right) \overset{\text{iid}}{\sim} \mathbb{P}_k^S$, where $i = 1, \ldots, B_k^S$ is the sample index. Meanwhile, for the target domain, we sample a batch of B^T examples as $X_i^T \overset{\text{iid}}{\sim} \mathbb{P}^T$, $i = 1, \ldots, B^T$.
- Let $f_\psi = h_\theta \circ g_\phi$ with $\psi = (\phi, \theta)$ be parameters of our deep net, wherein g_ϕ is the feature extractor and h_θ is the classifier on top of feature representations.

Constructing Challenging Samples

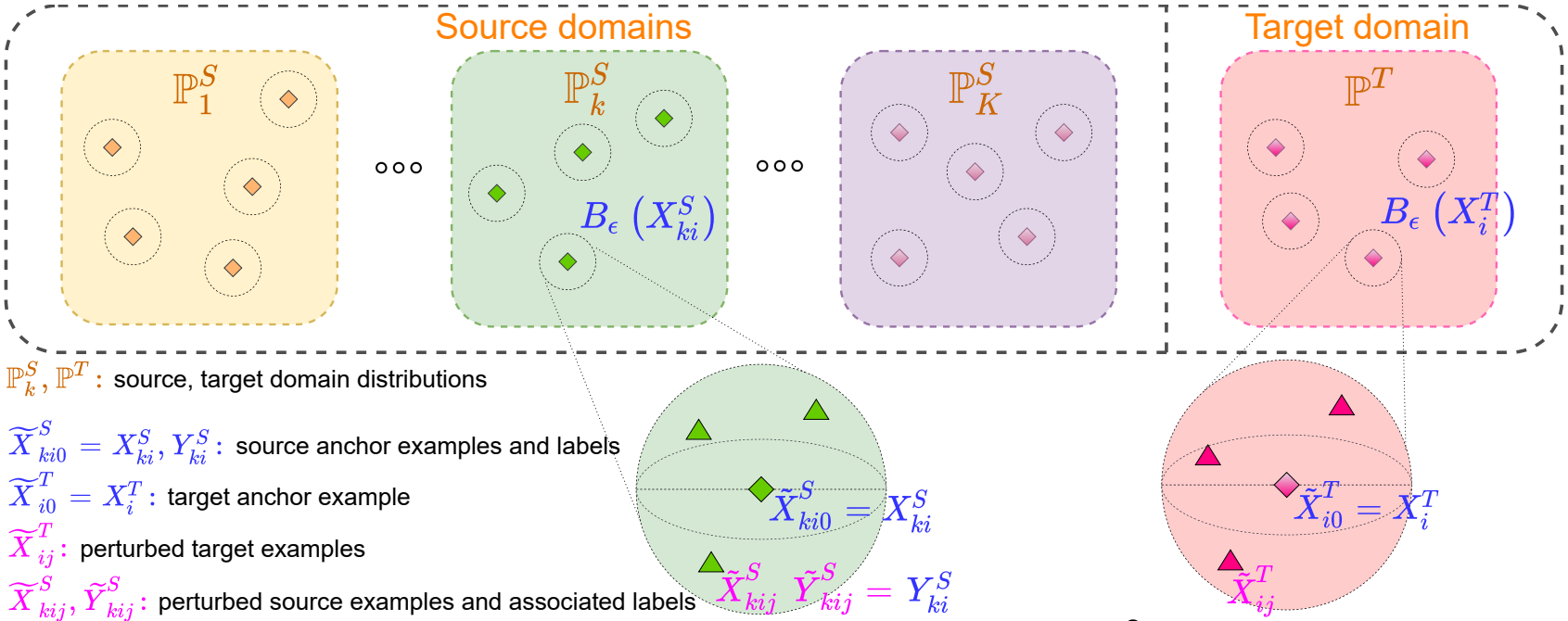

Figure 3: Overview of GLOT-DR. We sample $\left[X_{ki}^S, Y_{ki}^S\right]_{i=1}^{B_k^S}$ for each source domain, $\left[X_i^T\right]_{i=1}^{B^T}$ for the target domain, and define $Z, \tilde{Z}$ as in Eqs. (3,4). For $\left(Z, \tilde{Z}\right) \sim \gamma$ satisfying $\mathbb{E}_\gamma\left[\rho\left(Z, \tilde{Z}\right)\right]^{1/q} \leq \epsilon$, we have $\tilde{X}_{ki0}^S = X_{ki0}^S = X_{ki}^S$, $\tilde{X}_{i0}^T = X_{i0}^T = X_i^T$. Besides, $\tilde{X}_{kij}^S$ with $j \geq 1$ can be viewed as the *perturbed* examples in the ball $B_\epsilon\left(X_{ki}^S\right)$, which have the same label Y_{ki}^S. Similarly, $\tilde{X}_{ij}^T$ with $j \geq 1$ can be viewed as the *perturbed* examples in the ball $B_\epsilon\left(X_i^T\right)$.

Constructing Challenging Samples

Our method involves the construction of a random variable Z with distribution $\mathbb{P}$ and another random variable $\tilde{Z}$ with distribution $\tilde{\mathbb{P}}$, "containing" anchor samples $\left(X_{ki}^S, Y_{ki}^S\right), X_i^T$ and their perturbed counterparts $\left(\tilde{X}_{kij}^S, \tilde{Y}_{kij}^S\right), \tilde{X}_{ij}^T$

$$Z := \left[\left[\left[X_{kij}^S, Y_{kij}^S\right]_{k=1}^{K}\right]_{i=1}^{B_k^S}\right]_{j=0}^{n^S}, \left[\left[X_{ij}^T\right]_{i=1}^{B^T}\right]_{j=0}^{n^T}. \tag{3}$$

In contrast to Z, we next define random variable $\tilde{Z} \sim \tilde{\mathbb{P}}$, whose form is

$$\tilde{Z} := \left[\left[\left[\tilde{X}_{kij}^S, \tilde{Y}_{kij}^S\right]_{k=1}^{K}\right]_{i=1}^{B_k^S}\right]_{j=0}^{n^S}, \left[\left[\tilde{X}_{ij}^T\right]_{i=1}^{B^T}\right]_{j=0}^{n^T}. \tag{4}$$

Constructing Challenging Samples

The cost metric ρ is defined as:

$$\begin{aligned}\rho\left(Z,\tilde{Z}\right) := \infty \sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\left\|X_{ki0}^S - \tilde{X}_{ki0}^S\right\|_p^q + \infty \sum_{i=1}^{B^T}\left\|X_{i0}^T - \tilde{X}_{i0}^T\right\|_p^q \\ + \sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\sum_{j=1}^{n^S}\left\|X_{kij}^S - \tilde{X}_{kij}^S\right\|_p^q + \sum_{i=1}^{B^T}\sum_{j=1}^{n^T}\left\|X_{ij}^T - \tilde{X}_{ij}^T\right\|_p^q \\ + \infty \sum_{k=1}^{K}\sum_{i=1}^{B_k^S}\sum_{j=0}^{n^S}\rho_I\left(Y_{kij}^S, \tilde{Y}_{kij}^S\right),\end{aligned}$$

where ρ_I is a metric on the *label simplex* Δ_M.

Learning Robust Classifier

Upon clear definitions of $\tilde{Z}$ and $\tilde{\mathbb{P}}$, we wish to learn good representations and regularize the classifier f_ψ, via the following distributional robustness problem:

$$\min_{\theta,\phi} \max_{\tilde{\mathbb{P}}:\mathcal{W}_\rho\left(\mathbb{P},\tilde{\mathbb{P}}\right)\leq\epsilon} \mathbb{E}_{\tilde{Z}\sim\tilde{\mathbb{P}}}\left[r\left(\tilde{Z};\phi,\theta\right)\right]. \tag{5}$$

The cost function $r\left(\tilde{Z};\phi,\theta\right) := \alpha r^l\left(\tilde{Z};\phi,\theta\right) + \beta r^g\left(\tilde{Z};\phi,\theta\right) + \mathcal{L}\left(\tilde{Z};\phi,\theta\right)$ with $\alpha,\beta > 0$ is defined as the weighted sum of a *local-regularization function* $r^l\left(\tilde{Z};\phi,\theta\right)$, a *global-regularization function* $r^g\left(\tilde{Z};\phi,\theta\right)$, and the *loss function* $\mathcal{L}\left(\tilde{Z};\phi,\theta\right)$.

Learning Robust Classifier

Let us define
$\Gamma_\epsilon := \left\{ \gamma : \gamma \in \cup_{\tilde{\mathbb{P}}} \Gamma\left(\mathbb{P}, \tilde{\mathbb{P}}\right) \text{ and } \mathbb{E}_{(Z,\tilde{Z})\sim\gamma}\left[\rho\left(Z, \tilde{Z}\right)\right]^{1/q} \leq \epsilon \right\}$, and show that the inner max problem in Eq. (5) is equivalent to searching in Γ_ϵ.

Lemma 1

The optimization problem in Eq. (5) is equivalent to the following optimization problem:

$$\min_{\theta,\phi} \max_{\gamma:\in\Gamma_\epsilon} \mathbb{E}_{(Z,\tilde{Z})\sim\gamma}\left[r\left(\tilde{Z}; \phi, \theta\right)\right]. \tag{6}$$

Learning Robust Classifier

Concretely, we reach the following optimization problem with $\psi = (\phi, \theta)$:

$$\min_{\psi} \mathbb{E}_{\forall k:\left(X_{ki}^S, Y_{ki}^S\right)_{i=1}^{B_k^S} \overset{\text{iid}}{\sim} \mathbb{P}_k^S, X_{1:B^T}^T \overset{\text{iid}}{\sim} \mathbb{P}^T} \left[r\left(\tilde{Z}; \psi\right) \right], \tag{7}$$

where $r\left(\tilde{Z}; \psi\right)$ is defined as

$$\begin{aligned} &\mathbb{E}_{\left[\tilde{X}_{kij}^S\right]_j \sim q_{ki}^S} \left[\alpha s(X_{ki}^S, \tilde{X}_{kij}^S; \psi) + \ell(\tilde{X}_{kij}^S, Y_{ki}^S; \psi) \right] \\ &\quad + \mathbb{E}_{\left[\tilde{X}_{ij}^T\right]_j \sim q_i^T} \left[\alpha s\left(X_i^T, \tilde{X}_{ij}^T; \psi\right) \right] + \beta r^g \left(\left[X_{ki}^S\right]_{k,i}, \left[X_i^T\right]_i; \psi \right), \end{aligned} \tag{8}$$

and $s\left(\tilde{X}_0, \tilde{X}_j; \psi\right)$ measures the difference between 2 input samples.

Training Procedure of Our Approach

Algorithm 1: Projected Stein Variational Gradient Descent[4,5]

Input : A local distribution around X with an unnormalized density function $\tilde{p}(\cdot)$ and a set of initial particles $\{X^0\}_{i=1}^n$.

Output: A set of particles $\{X\}_{i=1}^n$ that approximates the local distribution corresponding to $\tilde{p}(\cdot)$.

for $l = 1$ *to* L **do**

$X^{l+1} = \prod_{B_\epsilon(X)} \left[X^l + \eta_l \hat{\phi}^*(X^l) \right]$

where $\hat{\phi}^*(X) = \frac{1}{n} \sum_{j=1}^n [k(X_j^l, X) \nabla_{X_j^l} \log \tilde{p}(X_j^l) + \nabla_{X_j^l} k(X_j^l, X)]$ and η_l is the step size at the l-th iteration.

return X^L

[4] Qiang Liu and Dilin Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Proceedings of NeurIPS*. vol. 29. 2016.

[5] Hoang Phan et al. "Stochastic Multiple Target Sampling Gradient Descent". In: *arXiv preprint arXiv:2206.01934* (2022).

Table of Contents

1. Introduction
2. Proposed approach
3. Experiments
4. Conclusion

Domain Generalization

Table 1: Single domain generalization accuracy (%) on CIFAR-10-C and CIFAR-100-C datasets with different backbone architectures. We use the **bold** font to highlight the best results.

Datasets	Backbone	Standard	Cutout	CutMix	AutoDA	Mixup	AdvTrain	ADA	ME-ADA	GLOT-DR
CIFAR-10-C	AllConvNet	69.2	67.1	68.7	70.8	75.4	71.9	73	78.2	**82.5**
	DenseNet	69.3	67.9	66.5	73.4	75.4	72.4	69.8	76.9	**83.6**
	WideResNet	73.1	73.2	72.9	76.1	77.7	73.8	79.7	83.3	**84.4**
	ResNeXt	72.5	71.1	70.5	75.8	77.4	73	78	83.4	**84.5**
	Average	71	69.8	69.7	74	76.5	72.8	75.1	80.5	**83.7**
CIFAR-100-C	AllConvNet	43.6	43.2	44	44.9	46.6	44	45.3	51.2	**54.8**
	DenseNet	40.7	40.4	40.8	46.1	44.6	44.8	45.2	47.8	**53.2**
	WideResNet	46.7	46.5	47.1	50.4	49.6	44.9	50.4	52.8	**56.5**
	ResNeXt	46.6	45.4	45.9	48.7	48.6	45.6	53.4	57.3	**58.4**
	Average	44.4	43.9	44.5	47.5	47.4	44.8	48.6	52.3	**55.7**

Domain Generalization

Table 2: Multi-source domain generalization accuracy (%) on PACS datasets.

	DSN	L-CNN	MLDG	Fusion	MetaReg	Epi-FCR	AGG	HEX	PAR	ADA	ME-ADA	GLOT-DR
Art	61.1	62.9	66.2	64.1	69.8	64.7	63.4	66.8	66.9	64.3	**67.1**	66.1
Cartoon	66.5	67.0	66.9	66.8	70.4	**72.3**	66.1	69.7	67.1	69.8	69.9	**72.3**
Photo	83.3	89.5	88.0	90.2	91.1	86.1	88.5	87.9	88.6	85.1	88.6	**90.4**
Sketch	58.6	57.5	59.0	60.1	59.2	65.0	56.6	56.3	62.6	60.4	63.0	**65.4**
Average	67.4	69.2	70.0	70.3	72.6	72.0	68.7	70.2	71.3	69.9	72.2	**73.5**

Table 3: Average classification accuracy (%) on MNIST benchmark.

	SVHN	MNIST-M	SYN	USPS	Average
Standard (ERM)	31.95± 1.91	55.96± 1.39	43.85± 1.27	79.92± 0.98	52.92± 0.98
PAR	36.08± 1.27	61.16± 0.21	45.48± 0.35	79.95± 1.18	55.67 ± 0.33
ADA	35.70 ± 2.00	58.65± 1.72	47.18± 0.61	80.40± 1.70	55.48± 0.74
ME-ADA	42.00± 1.74	63.98± 1.82	49.80± 1.74	79.10± 1.03	58.72± 1.12
GLOT-DR n=1	42.70 ± 1.03	67.72 ± 0.63	**50.53 ± 0.88**	82.32 ± 0.63	60.82 ± 0.79
GLOT-DR n=2	42.35 ± 1.44	67.95 ± 0.56	**50.53 ± 0.99**	82.33 ± 0.61	60.81± 0.90
GLOT-DR n=4	**43.10 ± 1.16**	**68.44 ± 0.46**	50.49 ± 1.04	**82.48 ± 0.51**	**61.13 ± 0.79**

Domain Adaptation

Table 4: Accuracy on Office-31.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DAN	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RTN	70.2	96.6	95.5	66.3	54.9	53.1	72.8
DANN	84.5	96.8	99.4	77.5	66.2	64.8	81.6
JAN	82	96.9	99.1	79.7	68.2	67.4	82.2
GTA	89.5	97.9	99.8	87.7	72.8	**71.4**	86.5
CDAN	93.1	98.2	**100**	89.8	70.1	68	86.6
DeepJDOT	88.9	98.5	99.6	88.2	**72.1**	70.1	86.2
ETD	92.1	**100**	**100**	88	71	67.8	86.2
GLOT-DR	**96.2**	98.9	**100**	**90.6**	69.9	69.6	**87.8**

Table 5: Accuracy on ImageCLEF-DA.

	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet	74.8	83.9	91.5	78.0	65.5	91.3	80.7
DAN	74.8	83.9	91.5	78.0	65.5	91.3	80.7
DANN	75.0	86.0	96.2	87.0	74.3	91.5	85.0
JAN	76.8	88.4	94.8	89.5	74.2	91.7	85.8
CDAN	76.7	90.6	97.0	90.5	74.5	93.5	87.1
ETD	**81.0**	91.7	97.9	**93.3**	79.5	95.0	89.7
GLOT-DR	**81.0**	**93.8**	**98.0**	**93.3**	**79.7**	**96.3**	**90.4**

Domain Adaptation

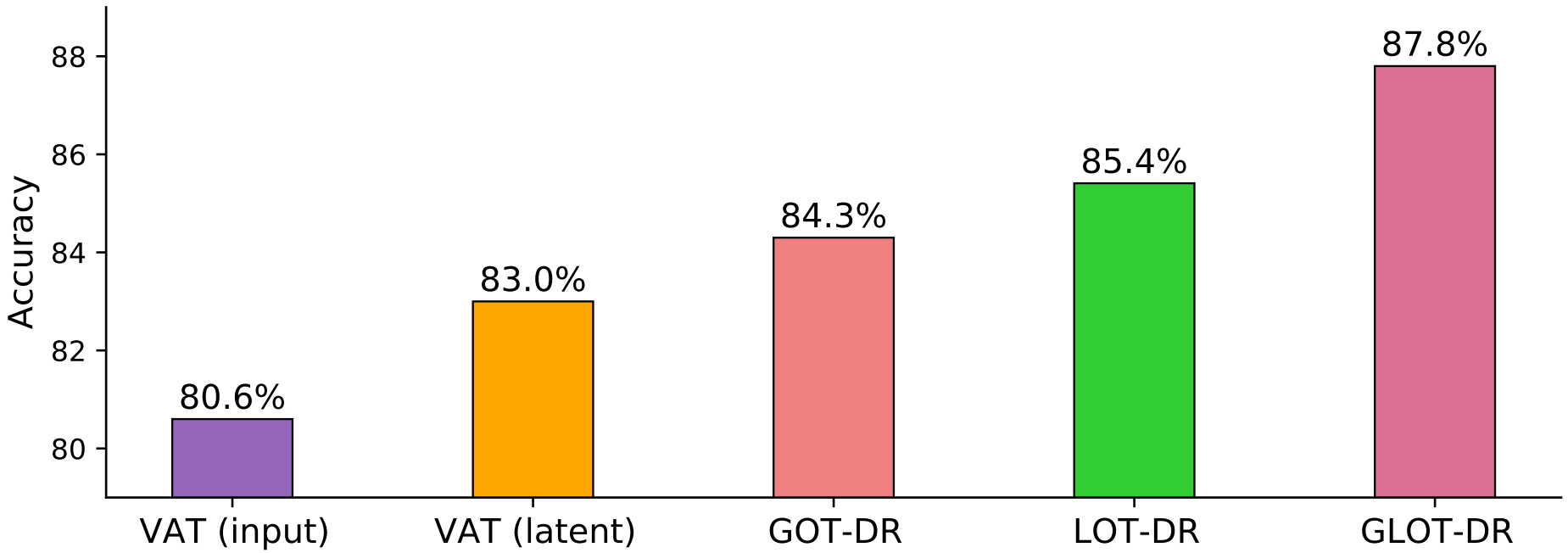

Figure 4: Average accuracy of ResNet50 on Office-31: Comparision between GLOT-DR's variants and VAT on the input and latent spaces.

Semi-supervised Learning

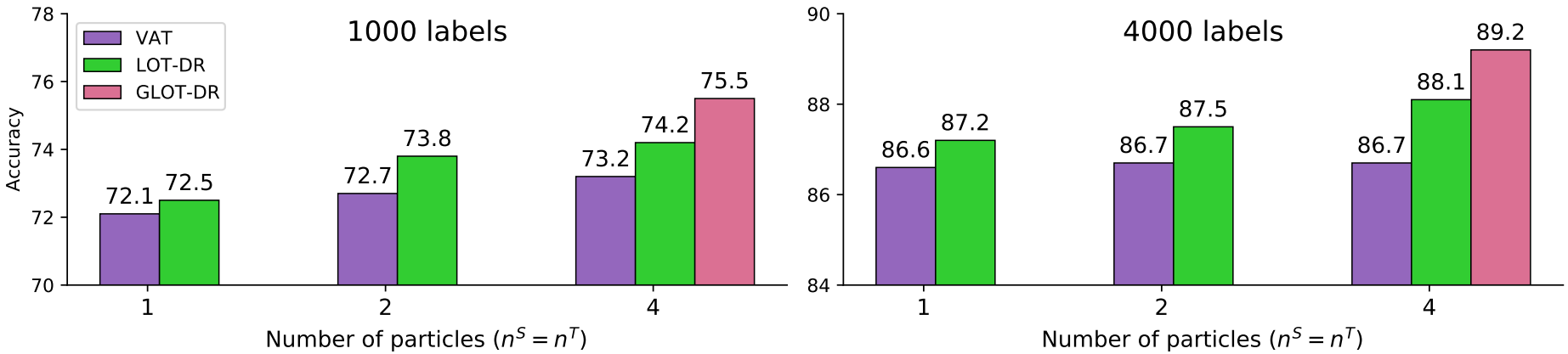

Figure 5: Accuracy (%) on CIFAR-10 of ConvLarge model in SSL settings when using 1,000 and 4,000 labeled examples (i.e. 100 and 400 labeled samples each class). Best viewed in color.

Semi-supervised Learning

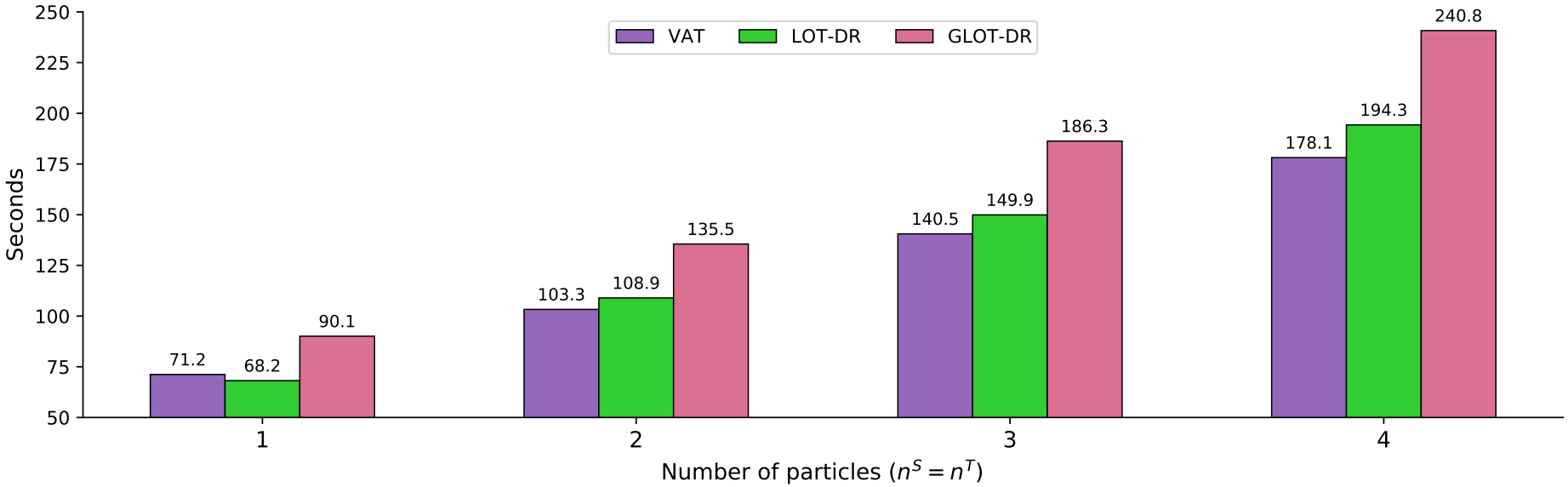

Figure 6: Running time of our proposed approach on: Intel(R) Xeon(R) CPU @ 2.00GHz CPU and Tesla P100 16GB VRAM GPU. Results are averaged over 3 runs.

Table 6: Adversarial robustness evaluation on CIFAR10 of ResNet18 model. PGD, AA and B&B represent the robust accuracy against the PGD attack (with 10/200 iterations), Auto-Attack and B&B attack, respectively, while NAT denotes the natural accuracy.

Method	NAT	PGD10	PGD200	AA	B&B
PGD-AT*	82.52	53.58	-	48.51	-
TRADES*	81.45	53.51	-	49.06	-
PGD-AT$^\diamond$	83.36	53.52	52.21	49.00	48.50
TRADES$^\diamond$	81.64	53.73	53.11	49.77	49.02
ADT-EXP	83.02	-	45.80	45.80	46.50
ADT-EXPAM	84.11	-	46.10	44.50	45.83
GLOT-DR	**84.13**	**54.13**	**53.18**	**49.94**	**49.40**

* Results are taken from Pang et al.

$^\diamond$ Our reproduced results.

Table of Contents

1 Introduction

2 Proposed approach

3 Experiments

4 Conclusion

Summary

- Although DR is a promising framework to improve neural network robustness and generalization capability, its current formulation shows some limitations, circumventing its application to real-world problems.
- In this thesis, we propose a rich OT based DR framework, named **G***lobal*-**L***ocal* **O***ptimal* **T***ransport based* **D***istributional* **R***obustness* (GLOT-DR) which is sufficiently rich for many real-world applications including DG, DA, SSL, and AML and has a closed-form solution.

Publications

Parts of this thesis are the extended/modified versions of the ones from the following papers:

- **Hoang Phan**, Ngoc Tran, Trung Le, Toan Tran, Nhat Ho, Dinh Phung. "Stochastic multiple target sampling gradient descent". Under review. *arXiv preprint arXiv:2206.01934*, 2022
- **Hoang Phan**, Trung Le, Trung Phung, Anh Bui, Nhat Ho, Dinh Phung. "Global-Local Regularization Via Distributional Robustness". Under review. *arXiv preprint arXiv:2203.00553*, 2022.

THANK YOU
FOR YOUR ATTENTION!