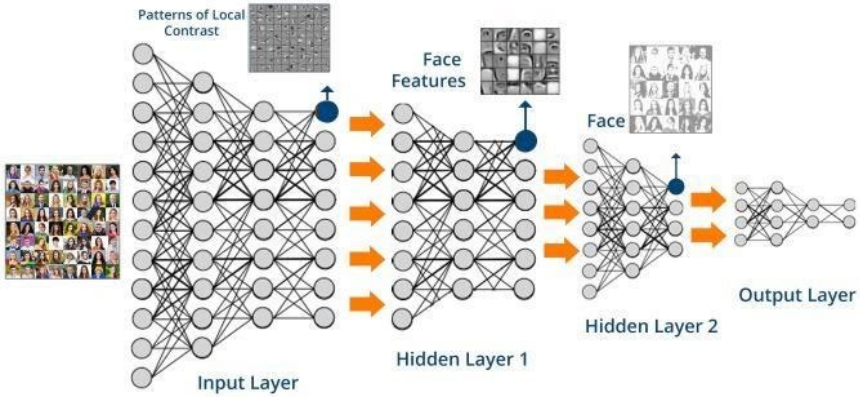
Deep Active Learning: Unified and Principled Method for Query and Training

Changjian Shui[1] **Fan Zhou**[1]
1 Université Laval

Christian Gagné[1] **Boyu Wang**[2]
2 University of Western Ontario

2019.12.4

Problem

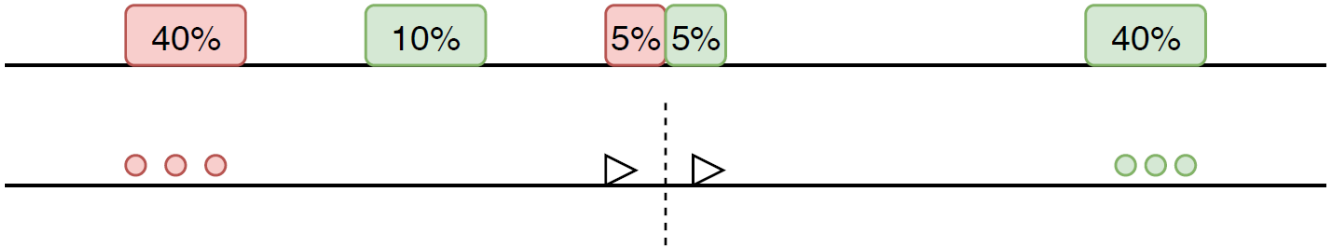


How to search the most informative samples in the context of DNN?

Apply DNN's output confidence score as an uncertainty acquisition function

Problem

A long term problem for uncertainty based sampling in AL is the so called sampling bias: **the current labeled points do not representative of the underlying distribution**.



01 Construct the core-sets through solving a K-center problem.

Problem

02 Heuristically selected a portion of samples according to the uncertainty score for exploitation and the rest portion used random sampling for exploration.

03 Collected samples whose gradients span a diverse set of directions for implicitly exploring these two.

In the context of deep AL, the available largescale unlabeled samples may be helpful to construct a good feature representation for potentially improving the performance.

Can we also additionally design the loss for DNN by leveraging the unlabeled samples during the Deep AL training ?

Active learning as distribution matching

$\hat{\mathcal{D}}$ are i.i.d generated by the underlying distribution $\mathcal{D}$

Labeling function $h^\star$ $\{(x_i, h^\star(x_i))\}_{i=1}^N$ $x_i \sim \mathcal{D}$

In the AL, the querying is not an i.i.d. procedure w.r.t. $\mathcal{D}$ the query procedure is an i.i.d. empirical process following a distribution $\mathcal{Q}$ with $\mathcal{Q} \neq \mathcal{D}$.

Interactive procedure can be viewed as estimating a proper $\mathcal{Q}$ to control the generalization error w.r.t. $(\mathcal{D}, h^\star)$.

Preliminaries

$$h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y} \quad \mathcal{X} \subseteq \mathbb{R}^d \quad \mathcal{Y} \in [0, 1]$$

loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$

expected risk $R_{\mathcal{D}}(h) = \mathbb{E}_{x \sim \mathcal{D}} \ell(h(x), h^\star(x))$

empirical risk $\hat{R}_{\mathcal{D}}(h) = \frac{1}{N} \sum_{i=1}^{N} \ell(h(x_i), y_i)$

$\mathbb{P}(\mathcal{X})$ is the set of all probability measures over $\mathcal{X}$

Assume that the loss ℓ is symmetric, L-Lipschtiz and M-upper bounded

$\forall h \in \mathcal{H}$ is at most H-Lipschtiz function.

Why Wasserstein distance

KL divergence

$$KL(P||Q) = \int_{-\infty}^{+\infty} p(x) log \frac{p(x)}{q(x)}$$

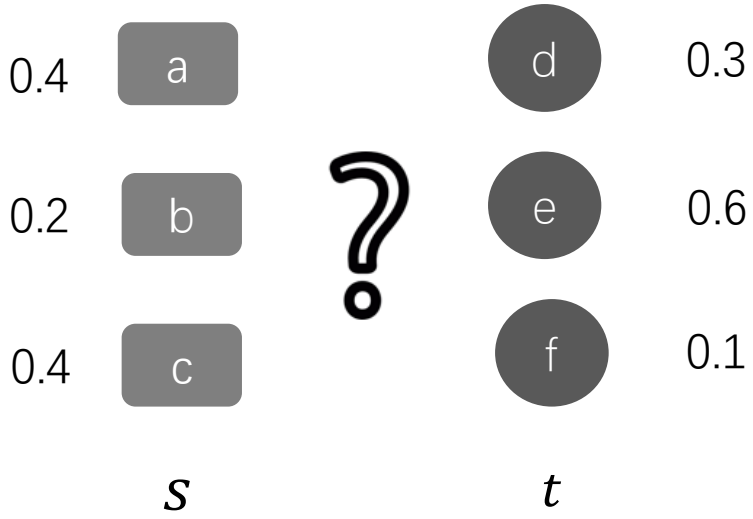
$p \sim N(0, \epsilon^3)$ $q \sim N(\epsilon, \epsilon^3)$ $KL(P||Q) = \frac{1}{2\epsilon}$

$$W(P_r, P_g) = \inf_{\gamma \in \prod(P_r, P_g)} \left(\int d(x,y)^p d\gamma(x,y) \right)^{1/p} = \inf_{\gamma \in \prod(P_r, P_g)} E_{(x,y)\sim\gamma}[\|x - y\|]$$

$\prod(P_r, P_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginal are respectively P_r and P_g. Intuitively, $\gamma(x, y)$ indicates how much "mass" must be transported from x to y in order to transform the distributions P_r into the distribution P_r. The EM distance is the "cost" of the optimal transport plan.

Why Wasserstein distance

Wasserstein distance



$d(s,t)$ $m(s,t)$

$$min \sum_{\pi(s,t)} d(s,t) * m(s,t)$$

$$\sum_{s} m(s,t) = b(t)\ \forall t$$

$$\sum_{t} m(s,t) = a(s)\ \ \forall s$$

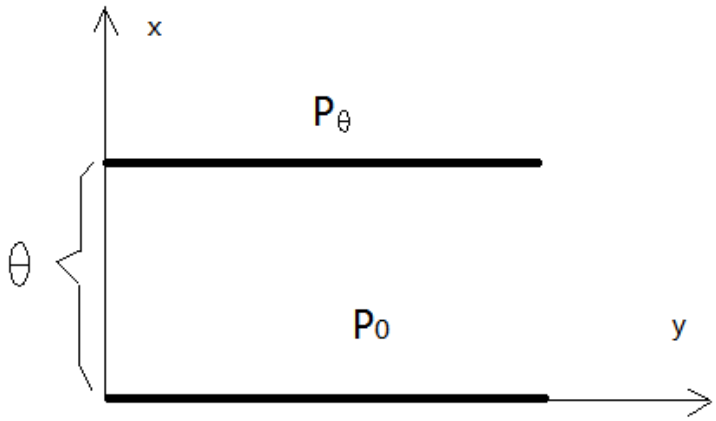
JS divergence $JS(p_r, p_g) = KL(p_r||p_m) + KL(p_g||p_m)$ $p_m = (p_r + p_g)$

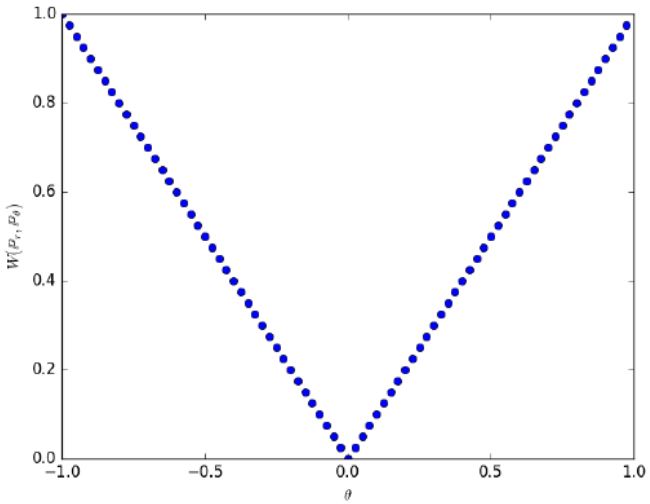
Why Wasserstein distance

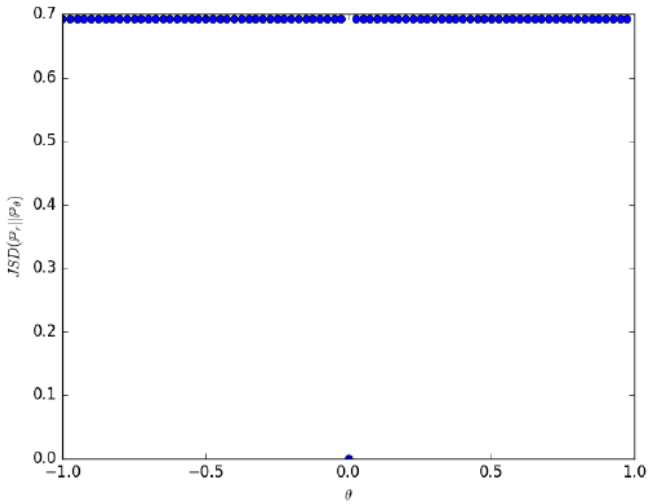
$$W(P_0, P_\theta) = |\theta|$$

$$JS(P_0, P_\theta) = \begin{cases} log2 & if\ \theta \neq 0 \\ 0 & if\ \theta = 0 \end{cases}$$

$$KL(P_0, P_\theta) = KL(P_\theta, P_0) = \begin{cases} +\infty & if \theta \neq 0 \\ 0 & if \theta = 0 \end{cases}$$







Why Wasserstein distance

$$W(P_r, P_g) = \inf_{\gamma \in \prod(P_r, P_g)} \left(\int d(x,y)^p d\gamma(x,y) \right)^{1/p}$$

Kantorovich-Rubinstein duality tells us that

$$W(P_r, P_\theta) = \sup_{||f||_{L \leq 1}} E_{x \sim P_r}[f(x)] - E_{x \sim p_\theta}[f(x)] \quad |f(x_1) - f(x_2)| \leq K|x_1 - x_2|$$

We replace $||f||_L \leq 1$ for $||f||_L \leq K$

$$W(P_r, P_\theta) = \frac{1}{K} \sup_{||f|| \leq K} E_{x \sim P_r}[f(x)] - E_{x \sim p_\theta}[f(x)]$$

$$KW(P_r, P_\theta) \approx \max_{w : |f_w|_L \leq K} E_{x \sim P_r}[f_w(x)] - E_{z \sim p(z)}[f_w(g_\theta(z))]$$

Why Wasserstein distance

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \left(\int d(x,y)^p d\gamma(x,y) \right)^{1/p}$$

$\mathcal{D} \in \mathbb{P}(\mathcal{X})$ $\mathcal{Q} \in \mathbb{P}(\mathcal{X})$

Optimal transport (or Monge-Kantorovich) problem can be defined as searching for a probabilistic coupling (joint probability distribution) $\gamma \in \mathbb{P}(\Omega \times \Omega)$ for $x_{\mathcal{D}} \sim \mathcal{D}$ $x_{\mathcal{Q}} \sim \mathcal{Q}$ that are minimizing the cost of transport w.r.t. some cost function c:

$$\begin{aligned} &\operatorname{argmin}_{\gamma} \int_{\mathcal{X} \times \mathcal{X}} c(x_{\mathcal{D}}, x_{\mathcal{Q}})^p d\gamma(x_{\mathcal{D}}, x_{\mathcal{Q}}), \\ &\text{s.t.} \quad \mathbf{P}^+ \# \gamma = \mathcal{D}; \quad \mathbf{P}^- \# \gamma = \mathcal{Q}, \end{aligned}$$

$\Pi(\mathcal{D}, \mathcal{Q})$ is the collection of all joint probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals $\mathcal{D}$ and $\mathcal{Q}$.

Why Wasserstein distance

$\mathcal{H}$-divergence

Variational adversarial active learning.

Bayesian generative active deep learning

4.1 The $\mathcal{H}$-divergence

Definition 1 (Based on Kifer et al. 2004) Given a domain $\mathcal{X}$ with $\mathcal{D}$ and $\mathcal{D}'$ probability distributions over $\mathcal{X}$, let $\mathcal{H}$ be a hypothesis class on $\mathcal{X}$ and denote by $I(h)$ the set for which $h \in \mathcal{H}$ is the characteristic function; that is, $\mathbf{x} \in I(h) \Leftrightarrow h(\mathbf{x}) = 1$. The *$\mathcal{H}$-divergence* between $\mathcal{D}$ and $\mathcal{D}'$ is

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{\mathcal{D}}[I(h)] - \Pr_{\mathcal{D}'}[I(h)] \right|.$$

H-divergence

所谓散度就是一个弱化的距离，他不一定具备距离的性质，比如有可能不满足对称性等等，那么所谓H是定义在假设空间$\mathcal{H}$的$\mathcal{D}$和$\mathcal{D}'$的距离：

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}}[h(x) = 1] - \Pr_{x \sim \mathcal{D}'}[h(x) = 1] \right|$$

直观来看，这个散度的意思是，在一个假设空间$\mathcal{H}$中，找到一个函数h，使得$\Pr_{x \sim \mathcal{D}}[h(x) = 1]$的概率尽可能大，而$\Pr_{x \sim \mathcal{D}'}[h(x) = 1]$的概率尽可能小，也就是说，我们用最大距离来衡量$\mathcal{D}, \mathcal{D}'$之间的距离。同时这个h也可以理解为是用来尽可能区分$\mathcal{D}, \mathcal{D}'$这两个分布的函数。

Why Wasserstein distance

H-divergence

所谓散度就是一个弱化的距离，他不一定具备距离的性质，比如有可能不满足对称性等等，那么所谓H是定义在假设空间$\mathcal{H}$的$\mathcal{D}$和$\mathcal{D}'$的距离：

$$d_{\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}}[h(x) = 1] - \Pr_{x \sim \mathcal{D}'}[h(x) = 1] \right|$$

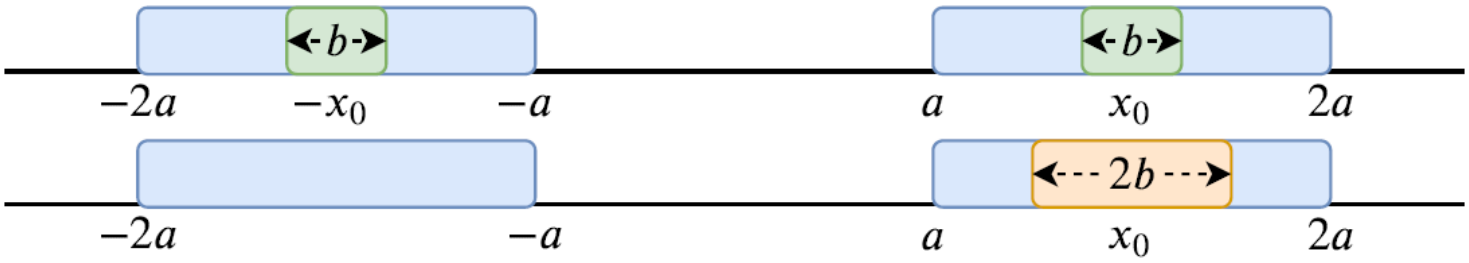
直观来看，这个散度的意思是，在一个假设空间$\mathcal{H}$中，找到一个函数h，使得$\Pr_{x \sim \mathcal{D}}[h(x) = 1]$的概率尽可能大，而$\Pr_{x \sim \mathcal{D}'}[h(x) = 1]$的概率尽可能小，也就是说，我们用最大距离来衡量$\mathcal{D}, \mathcal{D}'$之间的距离。同时这个h也可以理解为是用来尽可能区分$\mathcal{D}, \mathcal{D}'$这两个分布的函数。

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{Q}) = 2(1 - 2\epsilon)$$

- ϵ **prediction error** when training binary classifier to discriminate the observations sampling from these two distributions.

Thus smaller error means easy to separate the two distributions with larger H-divergence and vice versa.

Why Wasserstein distance



$$\mathcal{D}_1 \sim \mathcal{U}\big([-2a, -a] \cup [a, 2a]\big) \qquad \mathcal{D}_3 \sim \mathcal{U}\big([x_0 - b, x_0 + b]\big)$$

$$\mathcal{D}_2 \sim \mathcal{U}\big([-x_0 - \frac{b}{2}, -x_0 + \frac{b}{2}] \cup [x_0 - \frac{b}{2}, x_0 + \frac{b}{2}]\big)$$

$$\mathrm{supp}(\mathcal{D}_2) \subseteq \mathrm{supp}(\mathcal{D}_1) \qquad \mathrm{supp}(\mathcal{D}_3) \subseteq \mathrm{supp}(\mathcal{D}_1) \qquad a > b > 0$$

We set the classifier as decision stump $f(x) = \mathbf{1}\{x \geq p\}$

$$d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2) = d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_3) \qquad \min_{x_0} W_1(\mathcal{D}_1, \mathcal{D}_3) > \max_{x_0} W_1(\mathcal{D}_1, \mathcal{D}_2)$$

Labeling function assumption

Core-set	Assume the labeling function being Lipschtiz.
Plal:Cluster-based active learning	Probabilistic Lipschtiz condition.

Joint Probabilistic Lipschtiz property

Joint distribution optimal transportation for domain adaptation （NIPS2017）

Joint Probabilistic Lipschtiz property

Definition 1. *Let* $\phi : \mathbb{R} \to [0, 1]$. *We say labeling function* $h^\star$ *is* $\phi(\lambda)$-$(\mathcal{D}, \mathcal{Q})$ *Joint Probabilistic Lipschtiz if* $\mathrm{supp}(\mathcal{Q}) \subseteq \mathrm{supp}(\mathcal{D})$ *and for all* $\lambda > 0$ *and all distribution coupling* $\gamma \in \Pi(\mathcal{D}, \mathcal{Q})$*:*

$$\mathbb{P}_{(x_\mathcal{D}, x_\mathcal{Q}) \sim \gamma}[|h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q})| > \lambda \|x_\mathcal{D} - x_\mathcal{Q}\|_2] \leq \phi(\lambda) \tag{1}$$

Where $\phi(\lambda)$ *reflects the decay property. [14] showed that the faster the decay of* $\phi(\lambda)$ *with* $\lambda \to 0$*, the nicer of the distribution and the easier it is to learn the task.*

Joint Probabilistic Lipschtiz property

Joint distribution optimal transportation for domain adaptation（NIPS2017）

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\operatorname{argmin}} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) d\gamma(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2)$$

$$\mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) = \alpha \mathrm{d}(\mathbf{x}_1, \mathbf{x}_2) + \mathcal{L}(y_1, y_2)$$

$$\mathcal{P}_t^f = (\mathbf{x}, f(\mathbf{x}))_{\mathbf{x} \sim \mu_t} \qquad \hat{\mathcal{P}}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{\mathbf{x}_i^s, \mathbf{y}_i^s}$$

$$\hat{\mathcal{P}_t^f} = \frac{1}{N_t} \sum_{i=1}^{N_t} \delta_{\mathbf{x}_i^t, f(\mathbf{x}_i^t)}$$

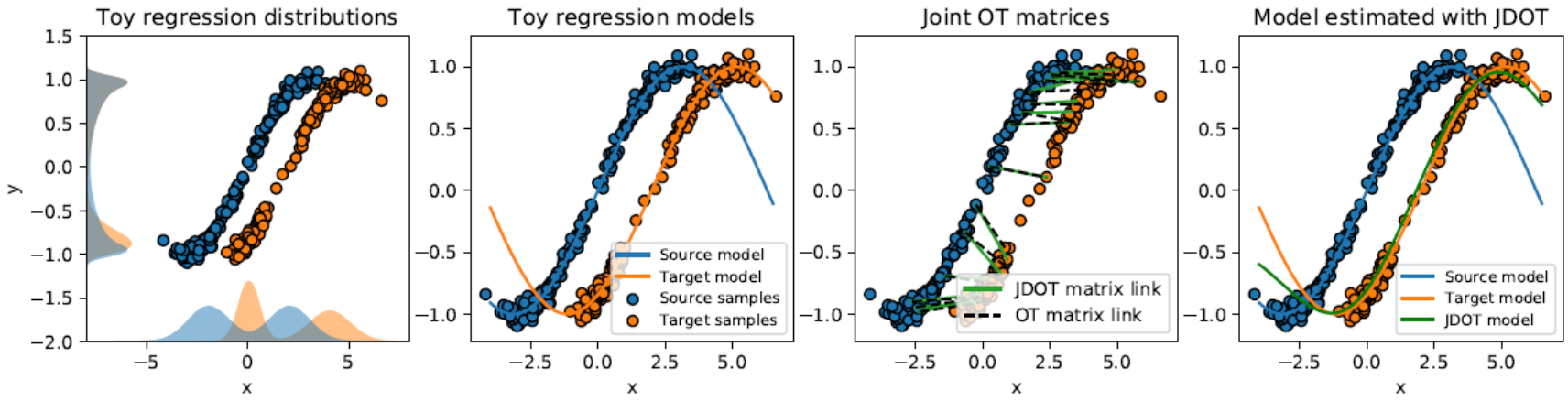
γ is then a matrix which belongs to Δ i.e. the transportation polytope of nonnegative matrices between uniform distributions.

Joint Probabilistic Lipschtiz property

Joint distribution optimal transportation for domain adaptation （NIPS2017）

Since our goal is to **estimate a prediction f on the target domain,** we propose to find the one **that produces predictions that match optimally source labels to the aligned target instances in the transport plan.**

$$\min_{f,\gamma\in\Delta} \sum_{ij} \mathcal{D}(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t))\gamma_{ij} \quad \equiv \quad \min_f W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f)$$



$$err_T(f) \stackrel{\text{def}}{=} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_t} \mathcal{L}(y, f(\mathbf{x}))$$

Bound related with querying distribution

Theorem 1. *Supposing $\mathcal{D}$ is the data generation distribution and $\mathcal{Q}$ is the querying distribution, if the loss ℓ is symmetric, L-Lipschtiz; $\forall h \in \mathcal{H}$ is at most H-Lipschtiz function and underlying labeling function $h^\star$ is $\phi(\lambda)$-$(\mathcal{D}, \mathcal{Q})$ Joint Probabilistic Lipschtiz, then the expected risk w.r.t. $\mathcal{D}$ can be upper bounded by:*

$$R_{\mathcal{D}}(h) \leq R_{\mathcal{Q}}(h) + L(H + \lambda)W_1(\mathcal{D}, \mathcal{Q}) + L\phi(\lambda)$$

Bound related with querying distribution

Corollary 1. *Supposing we have the finite observations which are i.i.d. generated from* $\mathcal{D}$ *and* $\mathcal{Q}$*:* $\hat{D} = \frac{1}{N}\sum_{i=1}^{N}\delta\{x_{\mathcal{D}}^{i}\}$ *and* $\hat{Q} = \frac{1}{N_q}\sum_{i=1}^{N_q}\delta\{x_{\mathcal{Q}}^{i}\}$ *with* $N_q \leq N$. *With probability* $\geq 1-\delta$*, the expected risk w.r.t.* $\mathcal{D}$ *can be further upper bounded by:*

$$R_{\mathcal{D}}(h) \leq \hat{R}_{\mathcal{Q}}(h) + L(H+\lambda)W_1(\hat{\mathcal{D}}, \hat{\mathcal{Q}}) + L\phi(\lambda) + 2L\mathrm{Rad}_{N_q}(h) + \kappa(\delta, N, N_q)$$

Practical deep batch active learning

$$\hat{L} = \frac{1}{L}\sum_{i=1}^{L}\delta\{x_i^l\} \quad \{y_i^l\}_{i=1}^{L} \qquad \hat{U} = \frac{1}{U}\sum_{i=1}^{U}\delta\{x_i^u\}$$

$$\hat{\mathcal{D}} = \hat{L} \cup \hat{U} \quad \text{with partial labels} \quad \{y_i^l\}_{i=1}^{L}$$

The goal of AL at each interaction is:

find a batch $\hat{B} = \frac{1}{B}\sum_{i=1}^{B}\delta\{x_i^b\}$ with $x_i^b \in \hat{U}$ during the query

find a hypothesis $h \in \mathcal{H}$

$$\min_{\hat{B},h} \hat{R}_{\hat{L}\cup\hat{B}}(h) + \mu W_1(\hat{\mathcal{D}}, \hat{L}\cup\hat{B})$$

Min-max Problem in DNN

https://mp.weixin.qq.com/s/Y8osoAfCKg9MInZ-q2Q33w

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right]$$

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

Kantorovich-Rubinstein duality

$$\min_{\hat{B},h} \hat{R}_{\hat{L}\cup\hat{B}}(h) + \mu W_1(\hat{\mathcal{D}}, \hat{L}\cup\hat{B})$$

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \hat{B}} \max_{\boldsymbol{\theta}^d} \hat{R}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^h) + \mu\hat{E}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^d)$$

$\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \boldsymbol{\theta}^d$ are parameters corresponding to the feature extractor, task predictor and critic;

$\hat{R}$ is the predictor loss

$\hat{E}$ is the adversarial (min-max) loss

Min-max Problem in DNN

$$\min_{\hat{B},h} \hat{R}_{\hat{L}\cup\hat{B}}(h) + \mu W_1(\hat{\mathcal{D}}, \hat{L}\cup\hat{B})$$

$$\min_{\boldsymbol{\theta}^f,\boldsymbol{\theta}^h,\hat{B}} \max_{\boldsymbol{\theta}^d} \hat{R}(\boldsymbol{\theta}^f,\boldsymbol{\theta}^h) + \mu\hat{E}(\boldsymbol{\theta}^f,\boldsymbol{\theta}^d)$$

$\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \boldsymbol{\theta}^d$ are parameters corresponding to the feature extractor, task predictor and critic;

$$h(x, y, (\boldsymbol{\theta}^f, \boldsymbol{\theta}^h)) \equiv h(x, y) : \mathcal{X} \times \mathcal{Y} \to (0, 1] \qquad \textstyle\sum_y h(x, y) = 1$$

parametric critic function $g(x, (\boldsymbol{\theta}^f, \boldsymbol{\theta}^d)) \equiv g(x) : \mathcal{X} \to [0, 1]$

$$\hat{R}(\boldsymbol{\theta}^f,\boldsymbol{\theta}^h) = \mathbb{E}_{(x,y)\sim\hat{L}\cup\hat{B}}\ell(h(x, y))$$

$$\hat{E}(\boldsymbol{\theta}^f,\boldsymbol{\theta}^d) = \mathbb{E}_{x\sim\hat{\mathcal{D}}}[g(x)] - \mathbb{E}_{x\sim\hat{L}\cup\hat{B}}[g(x)]$$

Min-max Problem in DNN

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \hat{B}} \max_{\boldsymbol{\theta}^d} \hat{R}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^h) + \mu \hat{E}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^d)$$

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \hat{B}} \max_{\boldsymbol{\theta}^d} \underbrace{\frac{1}{L+B} \sum_{(x,y)\in\hat{L}} \ell(h(x,y))}_{\text{Training: Prediction Loss}} + \underbrace{\mu\Big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - \big(\frac{1}{L+B} - \frac{1}{L+U}\big)\sum_{x\in\hat{L}} g(x)\Big)}_{\text{Training: Min-max Loss}}$$

$$+ \underbrace{\frac{1}{L+B} \sum_{(x,y^?)\in\hat{B}} \ell(h(x,y^?)) - \frac{\mu}{L+B}\sum_{x\in\hat{B}} g(x)}_{\text{Query}}$$

Training DNN

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h} \max_{\boldsymbol{\theta}^d} \frac{1}{L+B} \sum_{(x,y)\in\hat{L}} \ell(h(x,y)) + \mu\Big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - \big(\frac{1}{L+B} - \frac{1}{L+U}\big)\sum_{x\in\hat{L}} g(x)\Big)$$

Instead of only minimizing the prediction error, the proposed approach naturally **leveraged the information of unlabeled data** through a min-max training.

Training DNN

$$\min_{\boldsymbol{\theta}^f,\boldsymbol{\theta}^h,\hat{B}} \max_{\boldsymbol{\theta}^d} \hat{R}(\boldsymbol{\theta}^f,\boldsymbol{\theta}^h) + \mu\hat{E}(\boldsymbol{\theta}^f,\boldsymbol{\theta}^d)$$

$$\min_{\boldsymbol{\theta}^f,\boldsymbol{\theta}^h} \max_{\boldsymbol{\theta}^d} \frac{1}{L+B} \sum_{(x,y)\in\hat{L}} \ell(h(x,y)) + \mu\Big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - \big(\frac{1}{L+B} - \frac{1}{L+U}\big)\sum_{x\in\hat{L}} g(x)\Big)$$

The critic function g tried to evaluate how probable the sample comes from the labeled or unlabeled sets.

Given a fixed g , when $g(x) \rightarrow 1$ meaning the samples are highly probable from the unlabeled set and vice versa.

We should point out that it is the high level intuition,……

Since $B < U$ thus the $\frac{1}{L+B} - \frac{1}{L+U} > 0$ making this adversarial loss always valid.

cross entropy loss $l(x,y) = -\log h(x,y)$

Query strategy

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \hat{B}} \max_{\boldsymbol{\theta}^d} \hat{R}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^h) + \mu \hat{E}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^d)$$

$$\operatorname{argmin}_{\hat{B} \subset \hat{U}} \frac{1}{L+B} \sum_{(x, y^?) \in \hat{B}} \ell(h(x, y^?)) - \frac{\mu}{L+B} \sum_{x \in \hat{B}} g(x)$$

We do not known $y^?$ during the query

Optimize an upper bound of the loss.

Query strategy

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \hat{B}} \max_{\boldsymbol{\theta}^d} \hat{R}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^h) + \mu \hat{E}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^d)$$

$$\text{argmin}_{\hat{B} \subset \hat{U}} \frac{1}{L+B} \sum_{(x,y^?) \in \hat{B}} \ell(h(x, y^?)) - \frac{\mu}{L+B} \sum_{x \in \hat{B}} g(x)$$

Optimize an upper bound of the loss.

Cross entropy loss

$$y = \{1, \ldots, K\} \qquad \sum_{y \in \{1,\ldots,K\}} h(x, y) = 1$$

01

Minimizing over single worst case upper bound indicates the sample with highest least prediction confidence score:

$$\min_x \ell(h(x, y^?)) \leq \min_x \max_{y \in \{1,\ldots,K\}} -\log(h(x, y))$$

Query strategy

$$\min_x \ell(h(x, y^?)) \leq \min_x \max_{y \in \{1,\ldots,K\}} -\log(h(x, y))$$

$$h(x_1, \cdot) = [0.4, 0.6] \qquad h(x_2, \cdot) = [0.3, 0.7]$$

$$\max_y -\log(h(x_1, y)) < \max_y -\log(h(x_2, y))$$

02

Minimizing over ℓ_1 norm upper bound indicates the sample with uniformly of prediction confidence score:

$$\min_x \ell(h(x, y^?)) \leq \min_x \sum_{y \in \{1,\ldots,K\}} -\log(h(x, y))$$

Intuitively if the sample's prediction confidence is more uniform, the more uncertain of this sample will be.

Query strategy

Agnostic-label upper bound indicates uncertainty

Critic output indicates diversity

critic function $g(x) : \mathcal{X} \rightarrow [0, 1]$

If the critic function output trends to $g(x) \rightarrow 1$ it means $x \in \hat{U}$

According to the query loss, we want to sample the batch with higher critic values $g(x)$

meaning they look more different than the labelled samples under the Wasserstein distance.

Redundancy trick

$$\mu\Big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - \big(\frac{1}{L+B} - \frac{1}{L+U}\big)\sum_{x\in\hat{L}} g(x)\Big)$$

$$\gamma = \frac{U}{L} \qquad \alpha = \frac{B}{L}$$

$$\mu'\Big(\frac{1}{U}\sum_{x\in\hat{U}} g(x) - \frac{1}{\gamma}\frac{\gamma-\alpha}{1+\alpha}\frac{1}{L}\sum_{x\in\hat{L}} g(x)\Big) \qquad \mu' = \frac{\gamma}{1+\gamma}\mu$$

In optimizing the adversarial loss, we keep the same mini-batch size S for labelled and unlabeled observations.

$$\min_{\boldsymbol{\theta}^f}\max_{\boldsymbol{\theta}^d}\ \mu'\Big(\frac{1}{S}\sum_{x\in\hat{U}_S} g(x) - C_0\frac{1}{S}\sum_{x\in\hat{L}_S} g(x)\Big) \qquad C_0 = \frac{1}{\gamma^2}\frac{\gamma-\alpha}{1+\alpha}$$

WAAL

▷ **DNN Training stage**

for mini-batch of samples $\{(x^u)\}_{i=1}^{S}$ from $\hat{U}$ **do**

Constructing mini-batch $\{(x^l, y^l)\}_{i=1}^{S}$ from $\hat{L}$ through sampling with replacement (redundancy trick).

Updating $\boldsymbol{\theta}^h$: $\boldsymbol{\theta}^h = \boldsymbol{\theta}^h - \frac{\eta}{S} \sum_{(x^l,y^l)} \frac{\partial \ell(h((x^l,y^l))}{\partial \boldsymbol{\theta}^h}$

Updating $\boldsymbol{\theta}^f$: $\boldsymbol{\theta}^f = \boldsymbol{\theta}^f - \frac{\eta}{S}\big(\sum_{(x^l,y^l)} \frac{\partial \ell(h((x^l,y^l))}{\partial \boldsymbol{\theta}^f} + \mu' \{ \sum_{x^u} \frac{\partial g(x)}{\partial \boldsymbol{\theta}^f} - C_0 \sum_{x^l} \frac{\partial g(x)}{\partial \boldsymbol{\theta}^f} \} \big)$

Updating $\boldsymbol{\theta}^d$: $\boldsymbol{\theta}^d = \boldsymbol{\theta}^d + \frac{\eta\mu'}{S} \{ \sum_{x^u} \frac{\partial g(x)}{\partial \boldsymbol{\theta}^d} - C_0 \sum_{x^l} \frac{\partial g(x)}{\partial \boldsymbol{\theta}^d} \}$

end for

▷ **Querying stage**

Using a convex combination of Eq.(10),(11) to compute uncertainly score $\mathcal{U}(x^u)$, computing diversity score $g(x^u)$. Rank the score $\mathcal{U}(x^u) - \mu g(x^u)$ with $x^u \in \hat{U}$, choose the smallest B samples, forming querying batch $\hat{B}$

▷ **Updating**

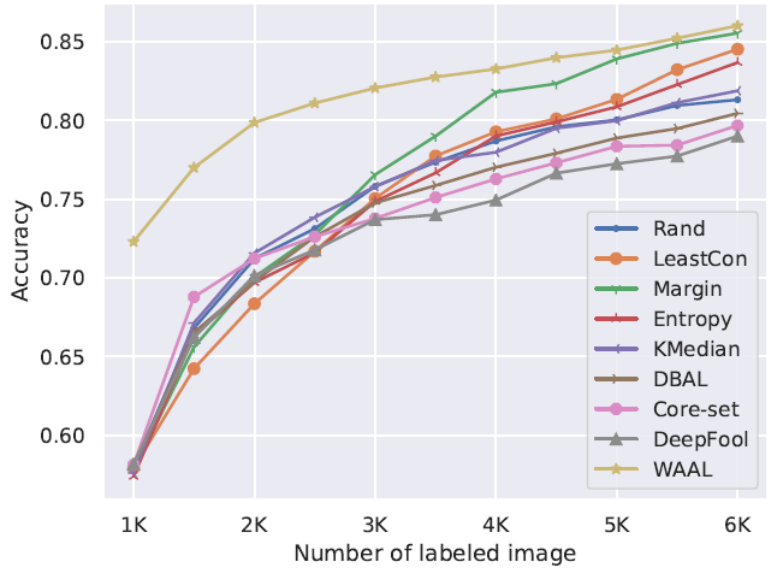
$\hat{L} = \hat{L} \cup \hat{B}, \hat{U} = \hat{U} \setminus \hat{B}$

Experiments

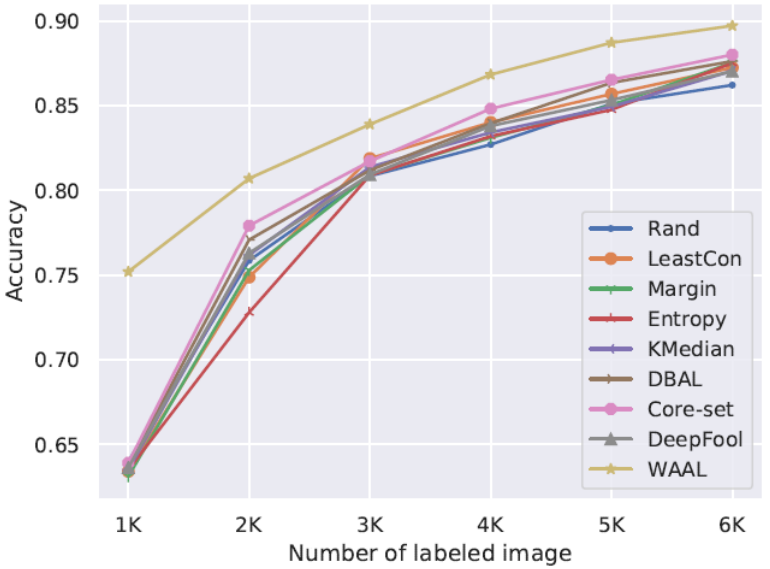
Feature extractor

On top of the feature extractor, we implement **2-layer multilayer** perceptron (MLP) as the **classifier and critic function**.

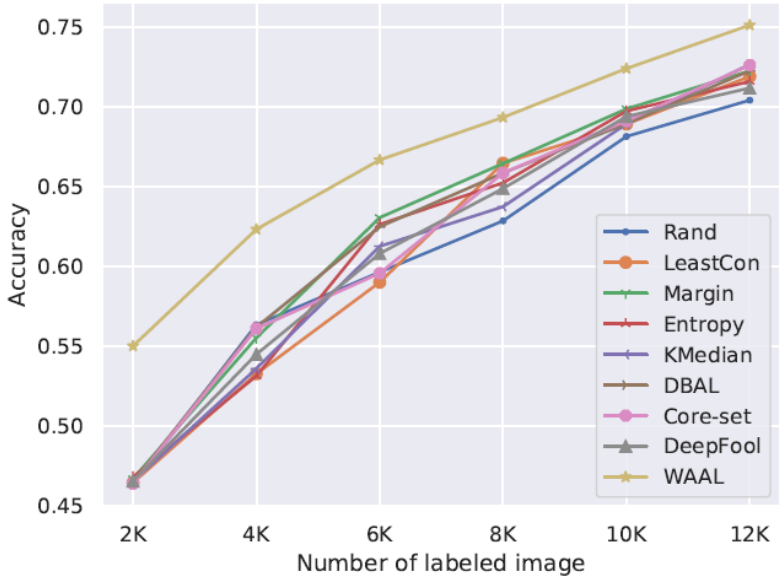
Method	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL



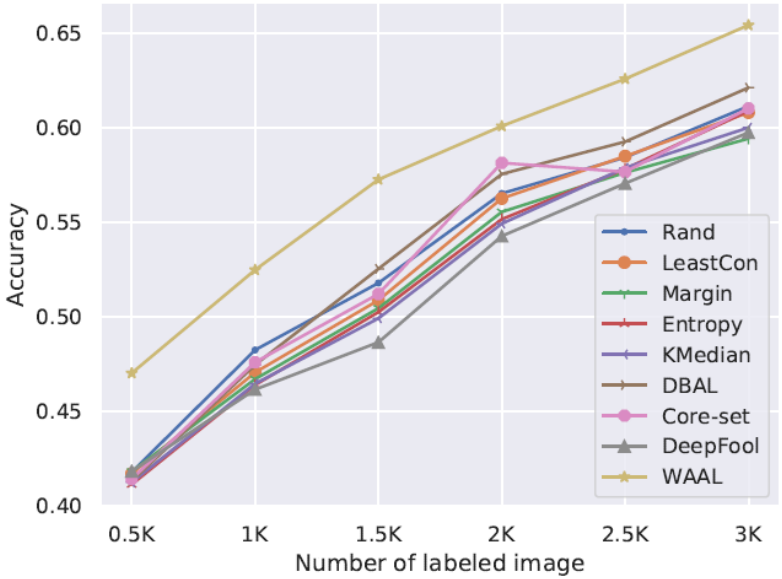
(a) FashionMNIST



(b) SVHN



(c) CIFAR10



(d) STL10

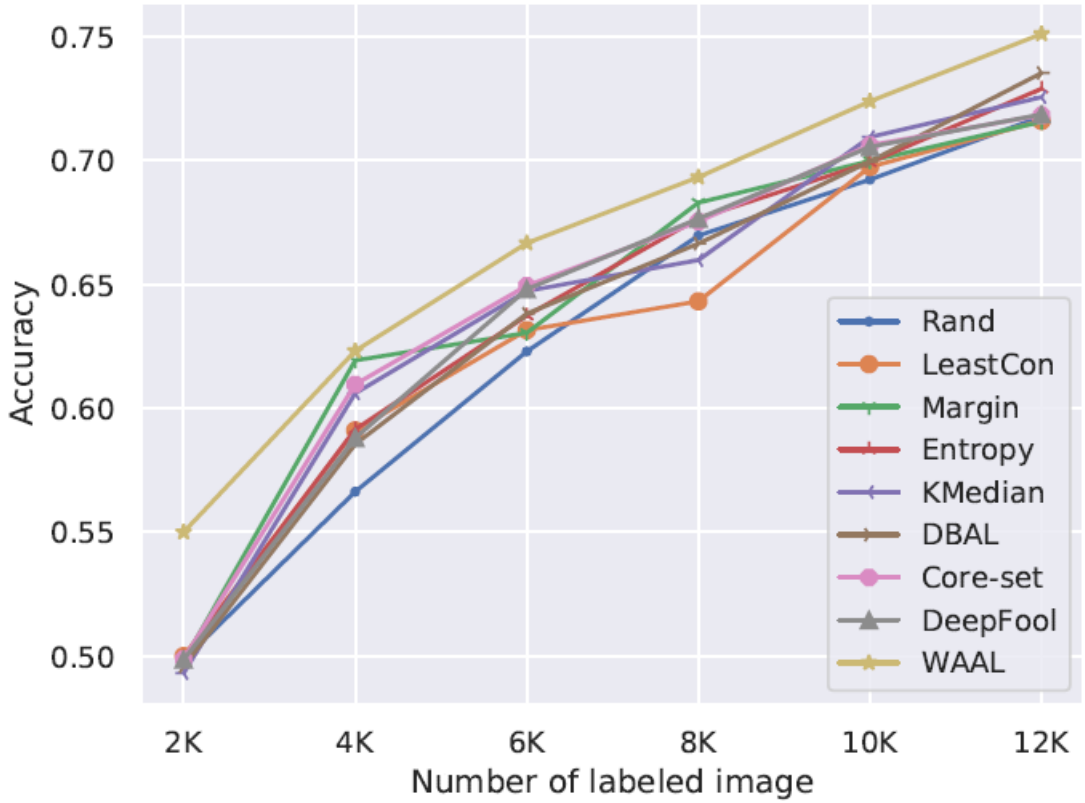


Figure 4: Ablation study in CIFAR10: the baselines are all trained by leveraging the unlabeled information through $\mathcal{H}$-divergence.

Conclusion

THANKS