Active Learning: Unified and Principled Method for Query and Training

Changjian Shui[1,⋆], Fan Zhou[1], Christian Gagné[1,2], Boyu Wang[3,4]

[⋆] changjian.shui.1@ulaval.ca; [1]Université Laval [2]Mila, Canada CIFAR AI Chair [3]University of Western Ontario [4]Vector Institute

Introduction

Active Learning (AL): Query the most informative samples to reduce the data annotation. Key factors in AL:

- *Query* How to select the most informative samples;
- *Training* How to train the model in the representation learning.

Contributions

A **unified** and **principled** approach for query and training in the deep AL.

- *Query* Explicit strategy with *Uncertain* and *Diverse* criteria
- *Training* Algorithm on the labeled and also leverage the unlabeled information.

Sampling Bias in AL

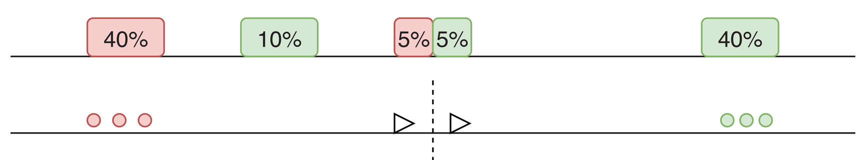


Figure 1: Sampling Bias in AL

AL as distribution matching

- Data distribution $\mathcal{D}(x)$, *Query* distribution $\mathcal{Q}(x)$
- Generalization error: $R_{\mathcal{D}}(h)$ with $R_{\mathcal{D}}(h) = \mathbb{E}_{x\sim\mathcal{D}}\ell(h(x), h^{\star}(x))$
- Metrics for measuring task similarity: Wasserstein-1 distance

Theorem 1 (Informal) *Supposing the transport cost in the Wasserstein distance is* $c(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$, *we have:*

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

Where the underlying labeling function $h^{\star}$ is $\phi(\lambda)$-$(\mathcal{D},\mathcal{Q})$ Joint Probabilistic Lipschitz

Why Wasserstein Distance

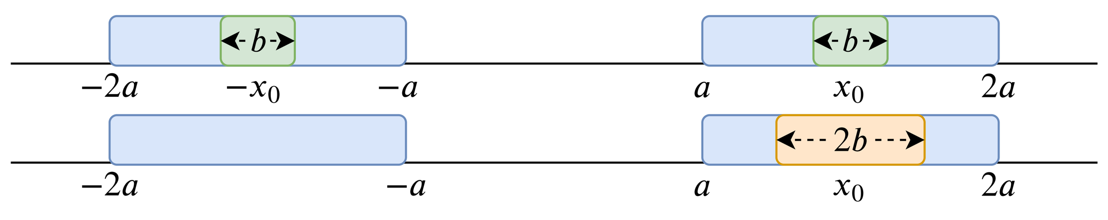


Figure 2: The desirable query distribution should be more diverse (first row) for avoiding *sampling bias* (second row).

Wasserstein distance exactly captures the property of diversity: *more diverse query distribution* $\mathcal{Q}$ *means smaller Wasserstein-1 distance* $W_1(\mathcal{D},\mathcal{Q})$.

Proposed Algorithm

Find a batch $\hat{B}$ and a hypothesis h to minimize:

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

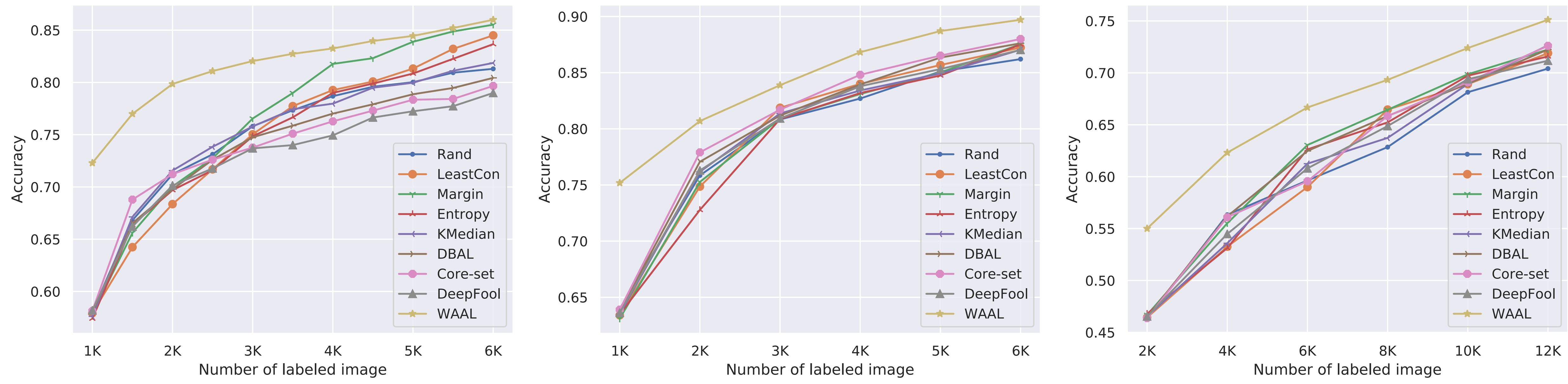
Introducing auxiliary task (dual term) to efficiently estimated W_1 distance.

- Neural networks parameters $\boldsymbol{\theta}^f$; $\boldsymbol{\theta}^h$; $\boldsymbol{\theta}^d$
- Alternative optimization over two training stages:

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

- Training procedure naturally leverages unlabeled data information.

Empirical Results



(a) Fmnist (b) SVHN (c) CIFAR10

Figure 3: Empirical performance on Fashion MNIST, SVHN and CIFAR-10 over five repetitions.

Method	LeastCon	Margin	Entropy	K-Median	DBAL	Core-set	DeepFool	WAAL
Time	0.94	0.95	0.95	33.98	9.25	45.88	124.46	1

Table 1: Relative Average querying time, assuming the query time of WAAL as the unit.

Query Strategies

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

where $y^?$ is the agnostic-label. Agnostic-label upper bound loss indicates uncertainty.

- Minimizing over the single worst case upper bound indicates the sample with the *highest least prediction confidence score*;
- Minimizing over ℓ_1 norm upper bound indicates the sample with a *uniformly of prediction confidence score*;

Critic output $g(x)$ indicates diversity.

Paper Link

Paper `https://arxiv.org/abs/1903.09109`
Code `https://github.com/cjshui/WAAL`

GANs v.s. Wasserstein

Similar task naturally extends the decision boundary of the original task.

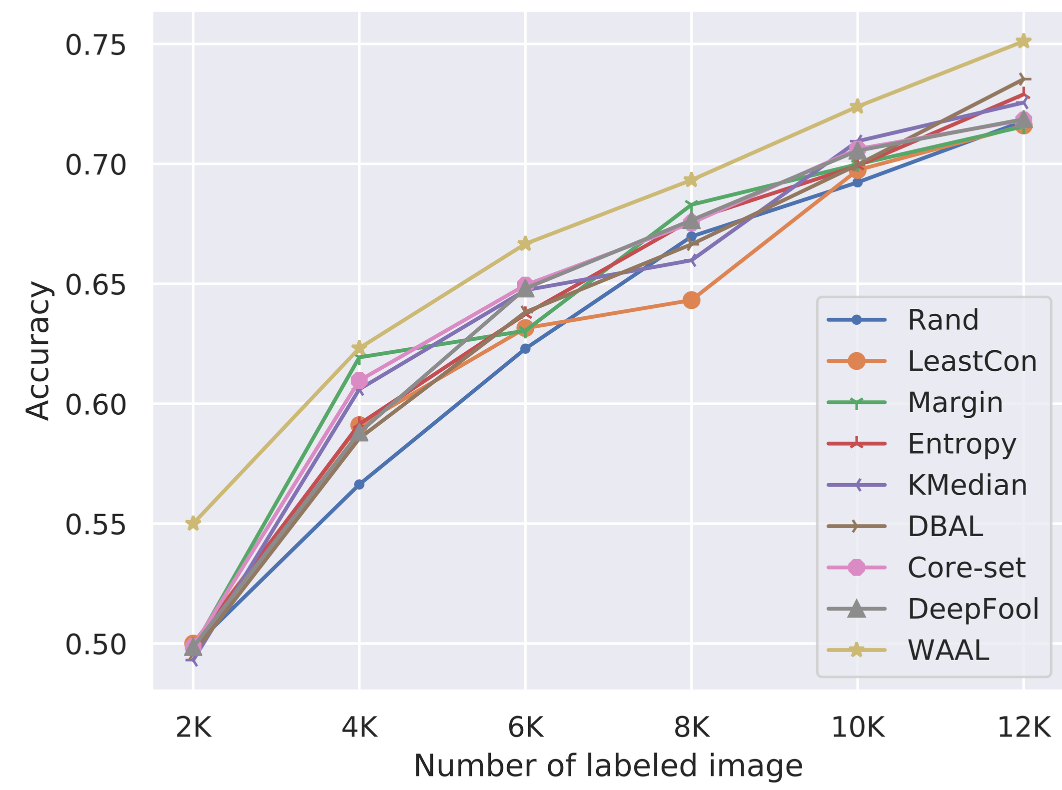


Figure 4: Ablation study in CIFAR-10: the baselines are all trained by leveraging the unlabeled information through $\mathcal{H}$-divergence.