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

![Sampling Bias in AL](Image)

Why Wasserstein Distance

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

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

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

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

- Neural networks parameters $\theta^f, \theta^h, \theta^d$
- Alternative optimization over two training stages:

$$\min_{w^f, w^h, w^d} \max_{\mu} \left[ \frac{1}{1 + B} \sum_{(x, x') \in L} \ell(h(x, y')) - \mu \sum_{x \in B} y(x) \right]$$

Training: Prediction Loss

$$\min_{\mu} \left[ \frac{1}{1 + B} \sum_{x \in B} \ell(h(x)) - \mu \sum_{x \in B} y(x) \right]$$

Query: Min-cost

$$\min_{\mu} \left[ \frac{1}{1 + B} \sum_{x \in B} \ell(h(x, y')) - \mu \sum_{x \in B} y(x) \right]$$

Training procedure naturally leverages unlabeled data information.

Proposed Algorithm

![Diagram of Proposed Algorithm](Image)

AL as distribution matching

- Data distribution $D(x)$, Query distribution $Q(x)$
- Generalization error: $R_D(h)$ with $R_D(h) = E_x, y \ell(h(x), y)$
- Metrics for measuring task similarity: Wasserstein-1 distance

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

$$R_D(h) \leq R_Q(h) + L(H + \lambda)W_1(D, Q) + Lo(\lambda).$$

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

Empirical Results

![Empirical Performance Graphs](Images)

Query Strategies

$$\arg \min_{h \in \mathcal{H}} L(h) = \frac{1}{1 + B} \sum_{x \in B} \ell(h(x), y') - \mu \sum_{x \in B} y(x)$$

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 $\ell_1$ norm upper bound indicates the sample with a uniformity of prediction confidence score;
- Critic output $g(x)$ indicates diversity.

GANs v.s. Wasserstein

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

Gans

![GANs Graph](Image)

Paper Link


Code [https://github.com/cjshui/WAAL](https://github.com/cjshui/WAAL)