NEURAL INFORMATION PROCESSING SYSTEMS

Enhancing Semi-Supervised Learning via Representative and Diverse Sample Selection

Qian Shao*, Jiangrui Kang*, Qiyuan Chen*, Zepeng Li, Hongxia Xu, Yiwen Cao, Jiajuan Liang†, Jian Wu†

Paper: https://arxiv.org/abs/2409.11653

Code: https://github.com/YanhuiAILab/RDSS

浙江大學 ZHEJIANG UNIVERSITY

微医 WE DOCTOR

北京师范大学 香港浸会大学 联合国际学院 BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY UNITED INTERNATIONAL COLLEGE

Backgrounds

The prevailing sample selection methods have many shortcomings.

Sampling methods in SSL:

- Random sampling may introduce imbalanced class distributions
- Stratified sampling is impractical in real-world scenarios
- Representativeness or diversity only sampling (see Fig. 1)

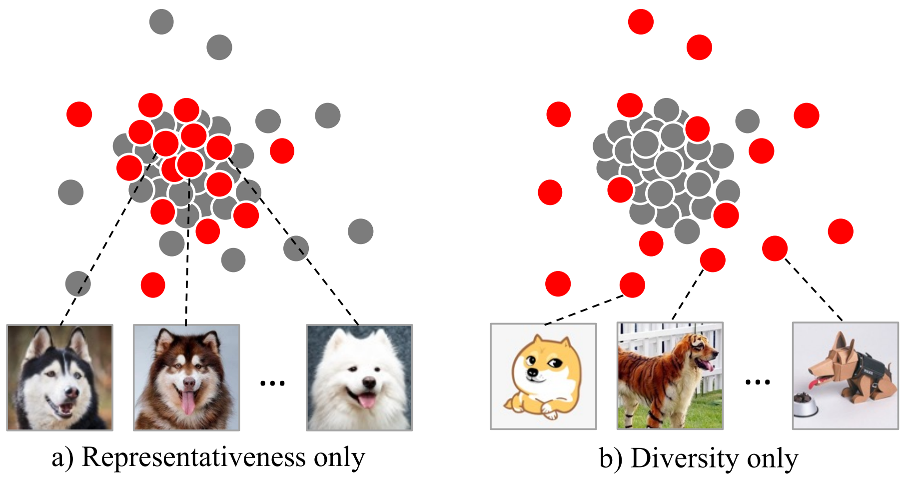

Fig. 1. Visualization of selected samples from a dog dataset using representativeness or diversity sampling methods.

Backgrounds

NEURAL INFORMATION PROCESSING SYSTEMS

Sampling methods in AL/SSAL:

- Begin with random samples
- Coupled with model training
- Human in the loop

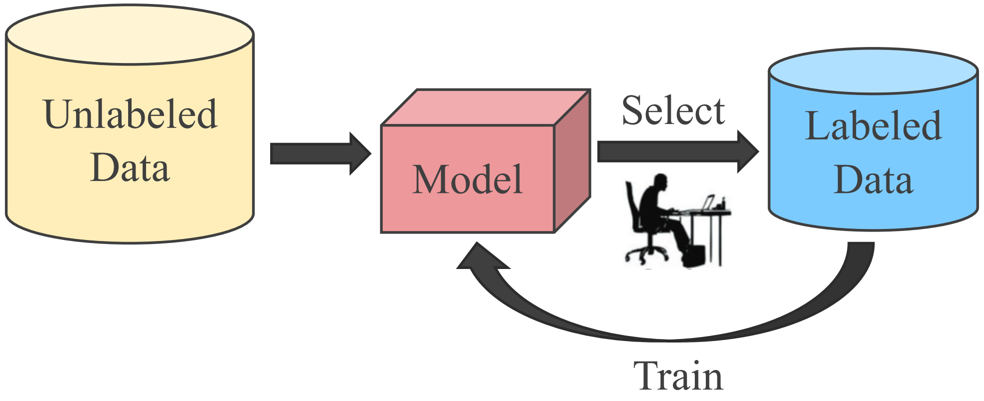

Fig. 2. AL-based sampling methods.

Methods

NEURAL INFORMATION PROCESSING SYSTEMS

Strategy: α-Maximum Mean Discrepancy

- Our goal can be formulated by solving: $\max_{\mathcal{J}_m \subset [n]} \mathrm{Rep}(X_{\mathcal{J}_m}, X_n) + \lambda \mathrm{Div}(X_{\mathcal{J}_m}, X_n)$,

  where Rep(·,·) and Div(·,·)quantify the representativeness and diversity of subdata respectively, and λ is a hyperparameter to balance the trade-off between representativeness and diversity.

- Quantification of representativeness and diversity

$$\mathrm{Rep}\left(X_{\mathcal{J}_m}, X_n\right) = -\mathrm{MMD}_k^2\left(X_{\mathcal{J}_m}, X_n\right)$$

$$= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k\left(\mathrm{x}_i, \mathrm{x}_j\right) - \frac{1}{m^2}\sum_{i\in\mathcal{J}_m}\sum_{j\in\mathcal{J}_m} k(\mathrm{x}_i, \mathrm{x}_j) + \frac{2}{mn}\sum_{i=1}^{n}\sum_{j\in\mathcal{J}_m} k(\mathrm{x}_i, \mathrm{x}_j), \quad (1)$$

$$\mathrm{Div}(X_{\mathcal{J}_m}, X_n) = -S_k\left(X_{\mathcal{J}_m}\right) = -\frac{1}{m^2}\sum_{i\in\mathcal{J}_m}\sum_{j\in\mathcal{J}_m} k\left(\mathrm{x}_i, \mathrm{x}_j\right), \quad (2)$$

where $k(\cdot,\cdot)$ is a kernel function on $\mathcal{X}\times\mathcal{X}$. Our optimization objective becomes:

$$\min_{\mathcal{J}_m \subset [n]} \mathrm{MMD}_k^2(X_{\mathcal{J}_m}, X_n) + \lambda S_k(X_{\mathcal{J}_m}). \quad (3)$$

Methods

NEURAL INFORMATION PROCESSING SYSTEMS

Strategy: α-Maximum Mean Discrepancy

Set $\lambda = \frac{1-\alpha}{\alpha m}$, since $\sum_{i=1}^{n} \sum_{j=1}^{n} k(\mathrm{x}_i, \mathrm{x}_j)$ is a constant, the objective function in (3) can be rewritten by

$$\alpha \mathrm{MMD}_k^2\left(X_{\mathcal{J}_m}, X_n\right) + \frac{1-\alpha}{m} S_k\left(X_{\mathcal{J}_m}\right) + \frac{\alpha(\alpha-1)}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k\left(\mathrm{x}_i, \mathrm{x}_j\right)$$

$$= \frac{\alpha^2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k\left(\mathrm{x}_i, \mathrm{x}_j\right) + \frac{1}{m^2} \sum_{i \in \mathcal{J}_m} \sum_{j \in \mathcal{J}_m} k(\mathrm{x}_i, \mathrm{x}_j) - \frac{2\alpha}{mn} \sum_{i=1}^{n} \sum_{j \in \mathcal{J}_m} k(\mathrm{x}_i, \mathrm{x}_j)$$

$$= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\frac{1}{m} \sum_{i \in \mathcal{J}_m} f(\mathrm{x}_i) - \frac{\alpha}{n} \sum_{j=1}^{n} f(\mathrm{x}_j) \right)^2, \qquad (4)$$

which defines a new concept called α-MMD, denoted by $\mathrm{MMD}_{k,\alpha}(X_{\mathcal{J}_m}, X_n)$.

Methods

NEURAL INFORMATION PROCESSING SYSTEMS

Algorithm: Modified Frank-Wolfe

- Theorem

With mild assumption on kernel and unlabeled data, $\min_{\mathcal{J}_m \subset [n]} \mathrm{MMD}_k^2(X_{\mathcal{J}_m}, X_n)$ can be solved by Frank-Wolfe algorithm with the following iterating formula:

$$\mathrm{x}_{i_{p+1}^*} \in \operatorname*{argmin}_{i \in [n]} f_{\mathcal{J}_p^*}(\mathrm{x}_i),\ \mathcal{J}_{p+1}^* \leftarrow \mathcal{J}_p^* \cup \{i_{p+1}^*\}, \mathcal{J}_0 = \emptyset, \qquad (5)$$

where $f_{\mathcal{J}_p^*}(\mathrm{x}_i) = \sum_{j \in \mathcal{J}_p} k(\mathrm{x}_i, \mathrm{x}_j) - \alpha p \sum_{l=1}^{n} k(\mathrm{x}_i, \mathrm{x}_l)$.

The corresponding algorithm of Eq. (5) may select repeated samples. To address this issue, we propose the Generalized Kernel Herding without Replacement (GKHR) algorithm based on Eq. (5):

$$\mathrm{x}_{i_{p+1}^*} \in \operatorname*{argmin}_{i \in [n] \setminus \mathcal{J}_p^*} f_{\mathcal{J}_p^*}(\mathrm{x}_i),\ \mathcal{J}_{p+1}^* \leftarrow \mathcal{J}_p^* \cup \{i_{p+1}^*\}, \mathcal{J}_0^* = \emptyset.$$

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Table 1. Comparison with other sampling methods, when applied to FlexMatch/FreeMatch.

Dataset	CIFAR-10			CIFAR-100			SVHN		STL-10	
Budget	40	250	4000	400	2500	10000	250	1000	40	250
Applied to FlexMatch [60]										
Stratified	91.45±3.41	95.10±0.25	95.63±0.24	50.23±0.41	67.38±0.45	73.61±0.43	89.60±1.86	93.66±0.49	75.33±3.74	92.29±0.64
Random	87.30±4.61	93.95±0.91	95.17±0.59	45.58±0.97	66.48±0.98	72.61±0.83	87.67±1.16	94.06±1.14	65.81±1.21	90.70±0.79
k-Means	81.23±8.71	94.59±0.51	95.09±0.65	41.60±1.24	65.99±0.57	71.53±0.42	90.28±0.69	93.82±1.04	55.43±0.39	90.64±1.05
USL [48]	91.73±0.13	94.89±0.20	95.43±0.15	46.89±0.46	66.75±0.37	72.53±0.32	90.03±0.63	93.10±0.78	75.65±0.60	90.77±0.36
ActiveFT [55]	70.87±4.14	93.85±1.37	95.31±0.75	25.69±0.64	57.19±2.06	70.96±0.75	89.32±1.87	92.53±0.43	55.57±1.42	87.28±1.19
RDSS (Ours)	**94.69±0.28**	**95.21±0.47**	**95.71±0.10**	**48.12±0.36**	**67.27±0.55**	**73.21±0.29**	**91.70±0.39**	**95.70±0.35**	**77.96±0.52**	**93.16±0.41**
Applied to FreeMatch [51]										
Stratified	95.05±0.15	95.40±0.23	95.80±0.29	51.29±0.56	67.69±0.58	73.90±0.53	92.58±1.05	94.22±0.78	79.16±5.01	91.36±0.18
Random	93.41±1.24	93.98±0.91	95.56±0.17	47.16±1.25	66.09±1.08	72.09±0.99	91.62±1.88	94.40±1.28	76.66±2.43	90.72±0.97
k-Means	88.05±5.07	94.80±0.48	95.51±0.37	44.07±1.94	66.09±0.39	71.69±0.72	93.30±0.46	94.68±0.72	63.22±4.92	89.99±0.87
USL [48]	93.81±0.62	95.19±0.18	95.78±0.29	47.07±0.78	66.92±0.33	72.59±0.36	93.36±0.53	94.44±0.44	76.95±0.86	90.58±0.58
ActiveFT [55]	78.13±2.87	94.54±0.81	95.33±0.53	26.67±0.46	56.23±0.85	71.20±0.68	92.60±0.51	93.71±0.54	63.31±2.99	86.60±0.30
RDSS (Ours)	**95.05±0.13**	**95.50±0.20**	**95.98±0.28**	**48.41±0.59**	**67.40±0.23**	**73.13±0.19**	**94.54±0.46**	**95.83±0.37**	**81.90±1.72**	**92.22±0.40**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Table 2. Comparison with AL approaches.

Dataset	CIFAR-10		CIFAR-100	
Budget	7500	10000	7500	10000
CoreSet	85.46	87.56	47.17	53.06
VAAL	86.82	88.97	47.02	53.99
LearnLoss	85.49	87.06	47.81	54.02
MCDAL	**87.24**	89.40	49.34	54.14
SL+RDSS (Ours)	87.18	**89.77**	**50.13**	**56.04**
Whole Dataset	95.62		78.83	

Table 3. Comparison with SSAL approaches.

Method	FlexMatch	FreeMatch
Stratified	91.45	95.05
Random	87.30	93.41
CoreSetSSL	87.66 ↑ 0.36	91.24 ↓ 2.17
MMA	74.61 ↓ 12.69	87.37 ↓ 6.04
CBSSAL	86.58 ↓ 0.72	91.68 ↓ 1.73
TOD-Semi	86.21 ↓ 1.09	90.77 ↓ 2.64
RDSS (Ours)	**94.69** ↑ 7.39	**95.05** ↑ 1.64

Table 4. Effect of different α.

Dataset	CIFAR-10			CIFAR-100		
Budget (m)	40	250	4000	400	2500	10000
0	85.54±0.48	93.55±0.34	94.58±0.27	39.26±0.52	63.77±0.26	71.90±0.17
0.40	92.28±0.24	93.68±0.13	94.95±0.12	42.56±0.47	65.88±0.24	71.71±0.29
0.80	94.42±0.49	94.94±0.37	95.15±0.35	45.62±0.35	66.87±0.20	72.45±0.23
0.90	94.33±0.28	95.03±0.21	95.20±0.42	48.12±0.50	67.14±0.16	72.15±0.23
0.95	94.44±0.64	95.07±0.26	95.45±0.38	**48.41±0.59**	67.11±0.29	72.80±0.35
0.98	94.51±0.39	95.02±0.15	95.31±0.44	48.33±0.54	**67.40±0.23**	72.68±0.22
1	94.53±0.42	95.01±0.23	95.54±0.25	48.18±0.36	67.20±0.29	73.05±0.18
$1-1/\sqrt{m}$ (Ours)	**95.05±0.13**	**95.50±0.20**	**95.98±0.28**	**48.41±0.59**	**67.40±0.23**	**73.13±0.19**

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks!

Paper

Code

浙江大學
ZHEJIANG UNIVERSITY

微医
WE DOCTOR

北京师范大学 香港浸会大学 联合国际学院
BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY
UNITED INTERNATIONAL COLLEGE