Meta-Query-Net: Resolving Purity-Informativeness Dilemma in Open-set Active Learning

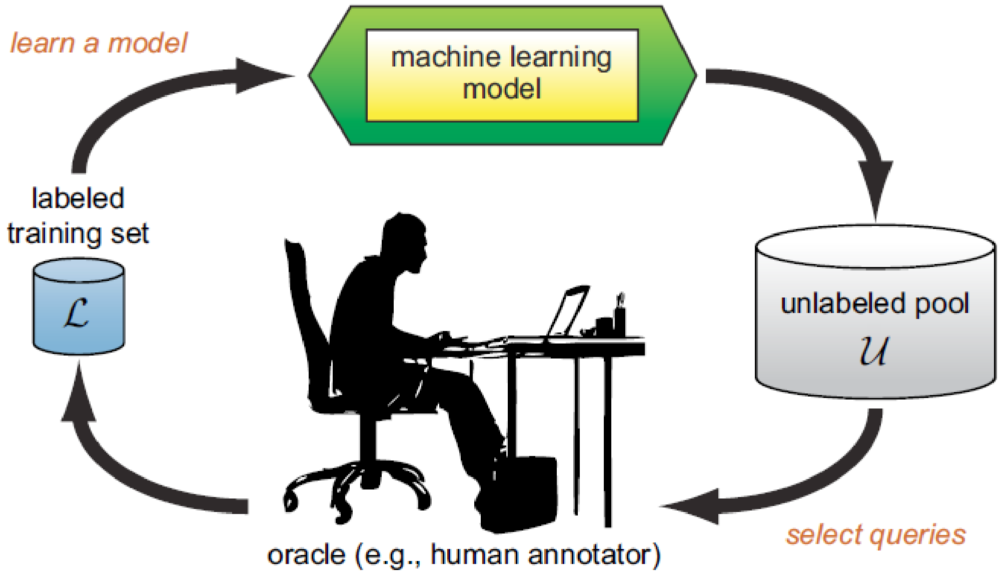
Dongmin Park (presenter), Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song*, Jae-Gil Lee*

Data Mining Lab, KAIST

Active Learning

Goal: Maximizing the model **performance** while **minimizing** labeling **costs**

→ Querying the examples that look maximally-informative



"Making an **AL algorithm** = Making a good **query strategy**"

Summary of Standard AL Approaches

$$x_{LC}^{*} = \underset{x}{\operatorname{argmax}}\, 1 - p_{\theta}(\hat{y}|x)$$

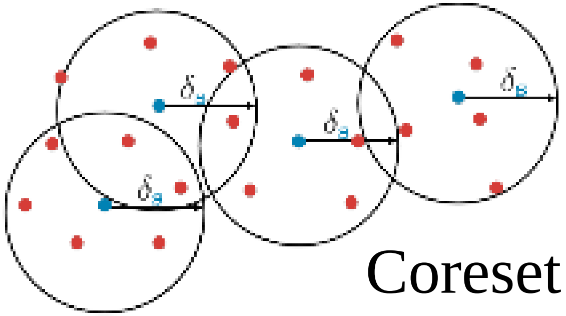
- **Uncertainty**-based
  - Querying the example that is **least certain** by the current model

    e.g., Softmax Confidence (CONF), Bayesian Disagreement (BALD), Learning Loss, …

- **Diversity**-based
  - Querying the example that **best represents** the entire data distribution

    e.g., Pre-clustering, Coreset, …

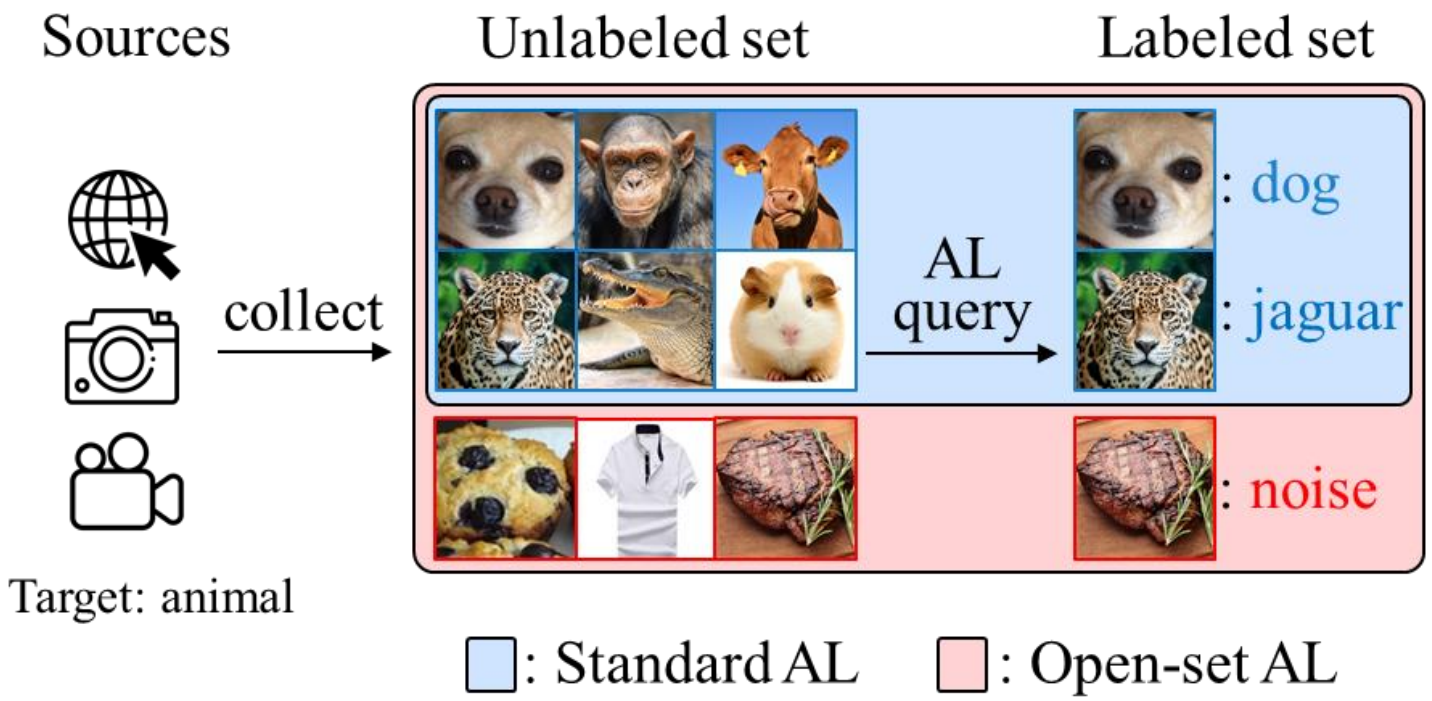
- **Hybrid**
  - BADGE, BatchBALD, …



Coreset

Open-set Active Learning: a more practical setup

- An unlabeled set consists of **only in-distribution** examples? → NO
  - Unlabeled data collected from **casual data curation** processes, *e.g.*, web-crawling, inevitably contains open-set noise, so called **out-of-distribution (OOD)** examples

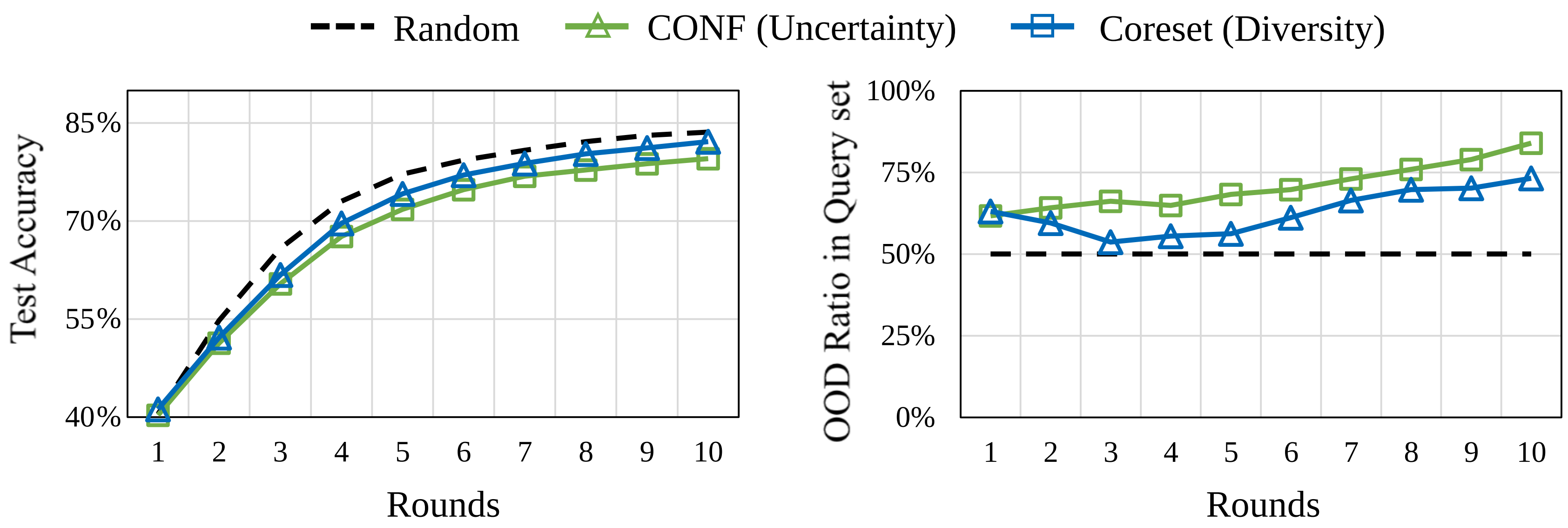


→ useless for a target task
= waste of the labeling budget!

Importance of Handling OOD in AL

- OOD examples are usually **uncertain & diverse**, thus often being queried
- This **wastes the labeling budget** and significantly **degrades AL performance**

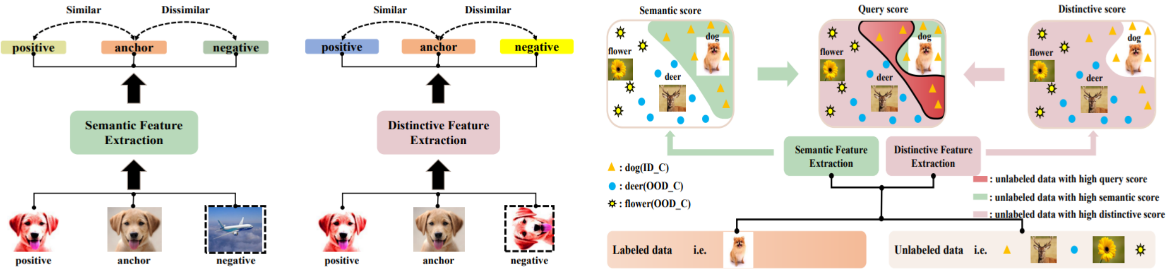
Datasets: [In: CIFAR10, OOD: SVHN] , Noise Ratio : 50%



→ Hinders the usability of AL in real-world applications!

Recent Open-set AL Approaches

- CCAL (ICCV'21)



(a) Contrast of semantics. (b) Contrast of distinctiveness. (c) Joint query strategy.

  - ➢ Learns two contrastive learners for calculating informativeness and OODness, respectively
  - ➢ Combines the two scores into a final query score using a **heuristic balancing rule**

- SIMILAR (NeurIPS'21)

SCMI	$I_f(\mathcal{A}; \mathcal{Q} \mid \mathcal{P})$
FLCMI	$\sum_{i \in \mathcal{U}} \max(\min(\max_{j \in \mathcal{A}} S_{ij}, \max_{j \in \mathcal{Q}} S_{ij}) - \max_{j \in \mathcal{P}} S_{ij}, 0)$
LogDetCMI	$\log \frac{\det(I - S_{\mathcal{P}}^{-1} S_{\mathcal{P},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{P},\mathcal{Q}}^T)}{\det(I - S_{\mathcal{A}\cup\mathcal{P}}^{-1} S_{\mathcal{A}\cup\mathcal{P},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{A}\cup\mathcal{P},\mathcal{Q}}^T)}$

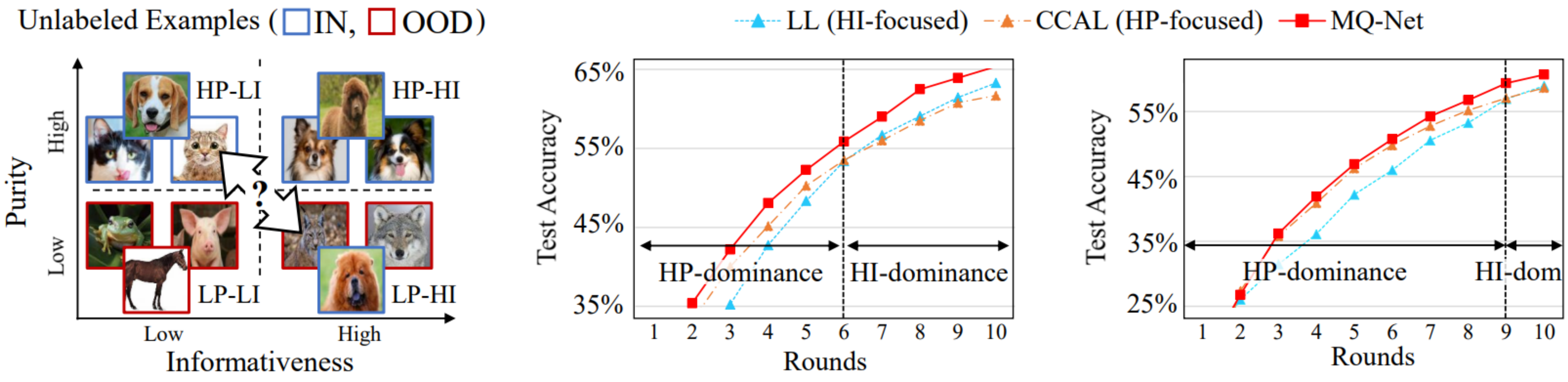
$$\max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} I_f(\mathcal{A}; \mathcal{I} \mid \mathcal{O})$$

  - ➢ Selects **a pure and core set** of examples by maximizing the distance coverage on the entire unlabeled data and jointly minimizing the distance coverage to the already labeled OOD data

→ Focus on **increasing purity of a query set** by effectively filtering out OOD examples

Purity-Informativeness Dilemma

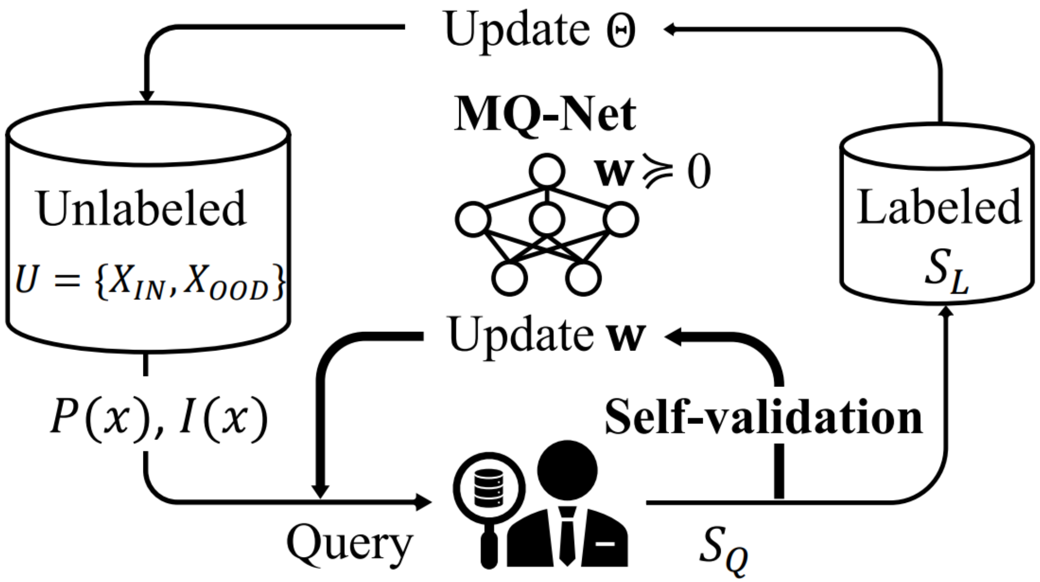
"Should we focus on the purity throughout the entire AL period?"



(a) Purity-informativeness Dilemma. (b) 10% Open-set Noise. (c) 30% Open-set Noise.

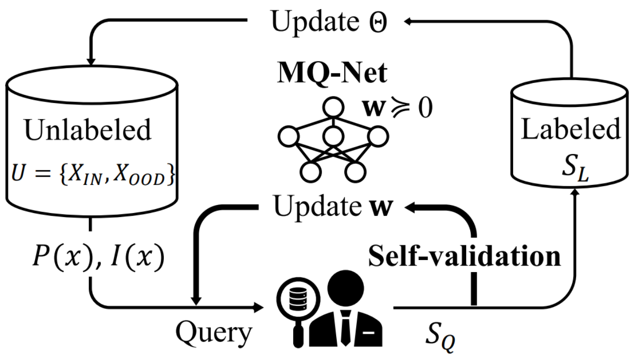
- Increasing Purity ↔ Losing Informativeness → Trade-off!
- The optimal trade-off changes according to AL rounds & noise ratios!

Meta-query-net (MQ-Net)



- To find the **best balance** between ***purity*** and ***informativeness***
- Learns a meta query-score function $\Phi(z_x;\ \boldsymbol{w})$
- Uses each round's **query set** as a **self-validation** set
- Can incorporate most existing AL scores and OOD scores

Objective of MQ-Net



IN>OOD High loss first

$$\ell_{mce}(x) = \mathbb{1}_{[l_x=1]}\ell_{ce}\big(f(x;\Theta), y\big),$$

$$\mathcal{L}(S_Q) = \sum_{i\in S_Q}\sum_{j\in S_Q} \max\Big(0, -\mathrm{Sign}\big(\ell_{mce}(x_i), \ell_{mce}(x_j)\big) \cdot \big(\Phi(z_{x_i};\mathbf{w}) - \Phi(z_{x_j};\mathbf{w}) + \eta\big)\Big)$$

$$\text{s.t. } \forall x_i, x_j, \text{ if } \mathcal{P}(x_i) > \mathcal{P}(x_j) \text{ and } \mathcal{I}(x_i) > \mathcal{I}(x_j), \text{ then } \Phi(z_{x_i};\mathbf{w}) > \Phi(z_{x_j};\mathbf{w}),$$

From a query set S_Q
(*unseen* for θ)

Skyline regularization
(to preserve order dominance w.r.t purity & informativeness)

- **Pairwise ranking loss** according to the masked cross entropy
- Output priority: 1) Informative IN examples first and 2) IN examples > OOD examples
- Stable optimization with ***skyline regularization***

Architecture of MQ-Net

Theorem 4.1. *For any MLP meta-model* $\mathbf{w}$ *with non-decreasing activation functions, a meta-score function* $\Phi(z;\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}$ *holds the skyline constraints if* $\mathbf{w} \succeq 0$ *and* $z(\in \mathbb{R}^d) \succeq 0$*, where* $\succeq$ *is the component-wise inequality.*

$$\Downarrow$$

$$\forall x_i, x_j, \;\; \text{if } \mathcal{P}(x_i) > \mathcal{P}(x_j) \text{ and } \mathcal{I}(x_i) > \mathcal{I}(x_j), \text{ then } \Phi(z_{x_i};\mathbf{w}) > \Phi(z_{x_j};\mathbf{w}),$$

- **Non-negative weights MLP**
  - Preserving order dominance between two examples w.r.t. purity and informativeness
  - Being attributed to the properties of non-decreasing activation functions
  - [Implementation] Applying a ReLU function for each parameters ***w*** (=differentiable)

→ Achieve the skyline constraint without any complex loss-based regularization!

Active Learning with MQ-Net

Meta-input Conversion

- Can incorporate any AL score $Q(x)$ and OOD score $O(x)$

- $P(x) = \text{Exp}(\text{Normalize}(-O(x)))$
- $I(x) = \text{Exp}(\text{Normalize}(Q(x)))$

Overall Procedure

Algorithm 1 AL Procedure with MQ-Net

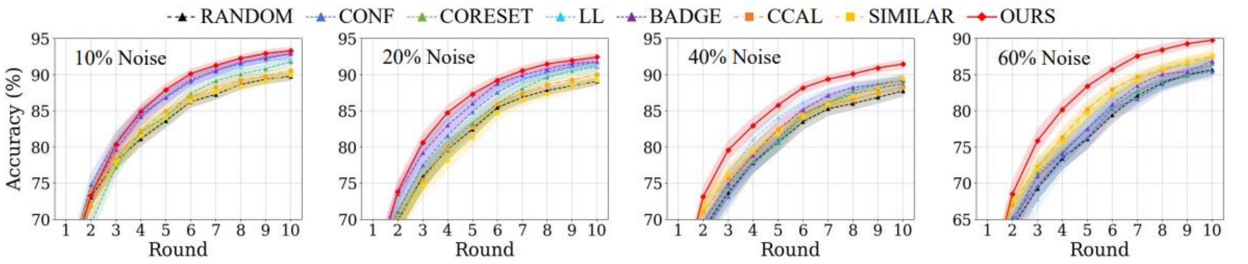
INPUT: S_L: labeled set, U: unlabeled set, r: number of rounds, Θ: parameters of target model, $\mathbf{w}$: parameters of MQ-Net

OUTPUT: Final target model Θ_*

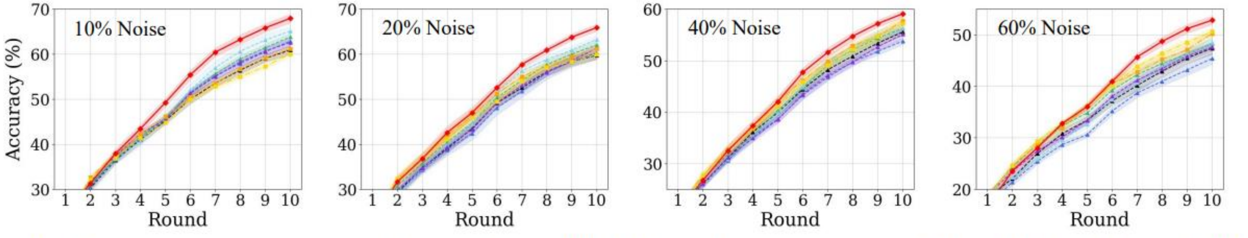
$\Theta_1, \mathbf{w}_1 \leftarrow$ Initialize the network parameters;
for $r = 1$ **to** r **do**
    /* Training the target model Θ*/
    $\Theta_* \leftarrow TrainingClassifier(S_L, \Theta_1)$
    /* Querying for the budget b */
    $S_Q \leftarrow \emptyset$;
    while $C(S_Q) \leq b$ **do**
        $S_Q \leftarrow S_Q \cup \arg\min(\Phi(U; \mathbf{w}))$;
    $S_L \leftarrow S_L \cup S_Q$; $U \leftarrow U - S_Q$;
    /* Training the meta-score function Φ */
    for $t = 1$ **to** meta-train-steps **do**
        Draw a mini-batch $\mathcal{M}$ and from S_Q;
        $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \alpha \nabla_{\mathbf{w}_t} \left(\mathcal{L}_{meta}(\mathcal{M})\right)$;
return Θ_*;

Experiments

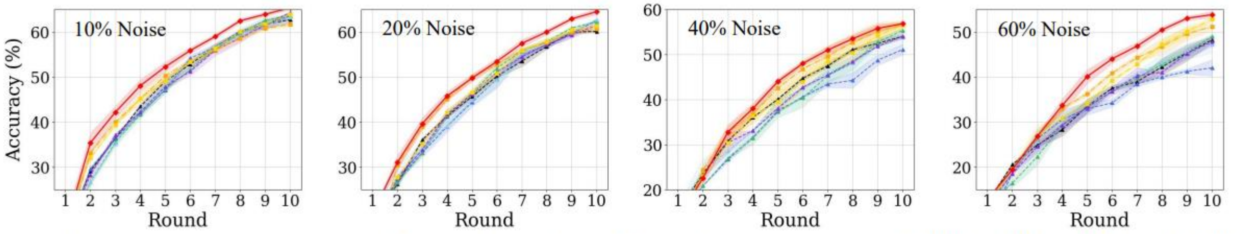
- On three datasets (CIFAR10, CIFAR100, ImageNet50) with varying noise ratios (10%, 20%, 40%, 60%)



(a) Accuracy comparison over AL rounds on CIFAR10 with open-set noise of 10%, 20%, 40%, and 60%



(b) Accuracy comparison over AL rounds on CIFAR100 with open-set noise of 10%, 20%, 40%, and 60%



(c) Accuracy comparison over AL rounds on ImgeNet with open-set noise of 10%, 20%, 40%, and 60%.

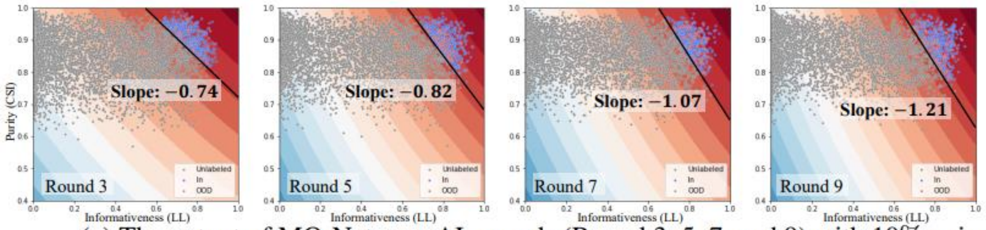
Table 1: Last test accuracy (%) at the final round for CIFAR10, CIFAR100, and ImageNet.

Datasets		CIFAR10 (4:6 split)				CIFAR100 (40:60 split)				ImageNet (50:950 split)			
Noise Ratio		10%	20%	40%	60%	10%	20%	40%	60%	10%	20%	40%	60%
Standard AL	CONF	92.83	91.72	88.69	85.43	62.84	60.20	53.74	45.38	63.56	62.56	51.08	45.04
	CORESET	91.76	91.06	89.12	86.50	63.79	62.02	56.21	48.33	63.64	62.24	55.32	49.04
	LL	92.09	91.21	89.41	86.95	65.08	64.04	56.27	48.49	63.28	61.56	55.68	47.3
	BADGE	92.80	91.73	89.27	86.83	62.54	61.28	55.07	47.60	64.84	61.48	54.04	47.80
Open-set AL	CCAL	90.55	89.99	88.87	87.49	61.20	61.16	56.70	50.20	61.68	60.70	56.60	51.16
	SIMILAR	89.92	89.19	88.53	87.38	60.07	59.89	56.13	50.61	63.92	61.40	56.48	52.84
Proposed	**MQ-Net**	**93.10**	**92.10**	**91.48**	**89.51**	**66.44**	**64.79**	**58.96**	**52.82**	**65.36**	**63.08**	**56.95**	**54.11**
% improve over 2nd best		0.32	0.40	2.32	2.32	2.09	1.17	3.99	4.37	0.80	1.35	0.62	2.40
% improve over the least		3.53	3.26	3.33	4.78	10.6	8.18	9.71	16.39	5.97	3.92	11.49	20.14

- MQ-Net achieves the **best accuracy** for all datasets
- MQ-Net is the most **robust** to any noise ratios
- In conclusion, MQ-Net finds the best trade-off between purity and informativeness

Takeaway & Ablation Studies

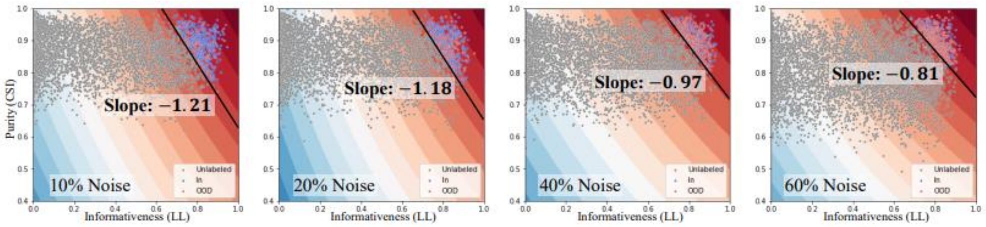
1. When the AL round progresses,



(a) The output of MQ-Net over AL rounds (Round 3, 5, 7, and 9) with 10% noise.

Purity (early) → Informativeness (late)

2. When the noise ratio increases,



(b) The final round's output of MQ-Net with varying noise ratios (10%, 20%, 40%, and 60%).

Informativeness (small) → Purity (high)

Table 2: Effect of the meta inputs to MQ-Net.

Dataset		CIFAR10 (4:6 split)			
Noise Ratio		10%	20%	40%	60%
Standard AL	BADGE	92.80	91.73	89.27	86.83
Open-set AL	CCAL	90.55	89.99	88.87	87.49
MQ-Net	CONF-ReAct	93.21	91.89	89.54	87.99
	CONF-CSI	**93.28**	**92.40**	91.43	89.37
	LL-ReAct	92.34	91.85	90.08	88.41
	LL-CSI	93.10	92.10	**91.48**	**89.51**

Table 3: Efficacy of the self-validation set.

Dataset		CIFAR10 (4:6 split)			
Noise Ratio		10%	20%	40%	60%
MQ-Net	Query set	**93.10**	**92.10**	**91.48**	**89.51**
	Random	92.10	91.75	90.88	87.65

Table 4: Efficacy of the skyline constraint.

Noise Ratio		10%	20%	40%	60%
MQ-Net	w/ skyline	**93.10**	**92.10**	**91.48**	**89.51**
	w/o skyline	87.25	86.29	83.61	81.67

THANK YOU
Any Question?