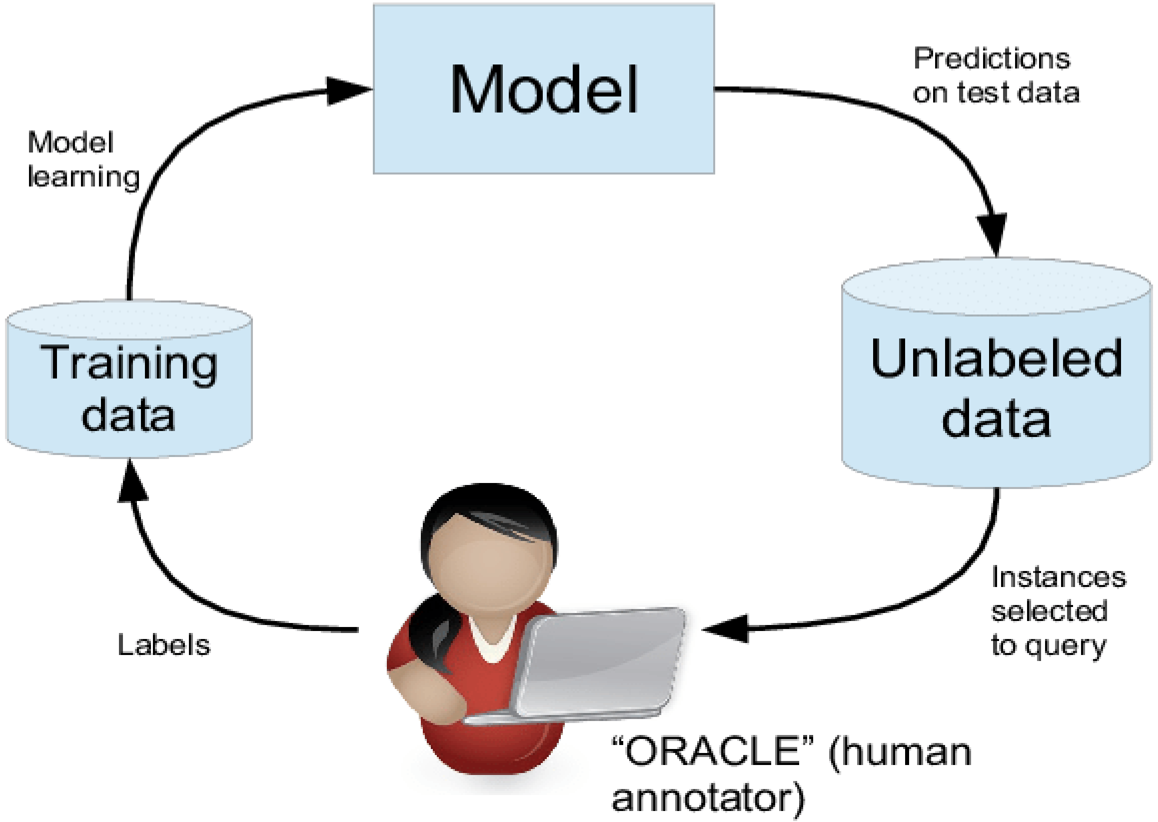
SIMILAR
Submodular Information Measures Based Active Learning In Realistic Scenarios

Suraj Kothawade*, Nathan Beck, Krishnateja Killamsetty, Rishabh Iyer

*suraj.kothawade@utdallas.edu

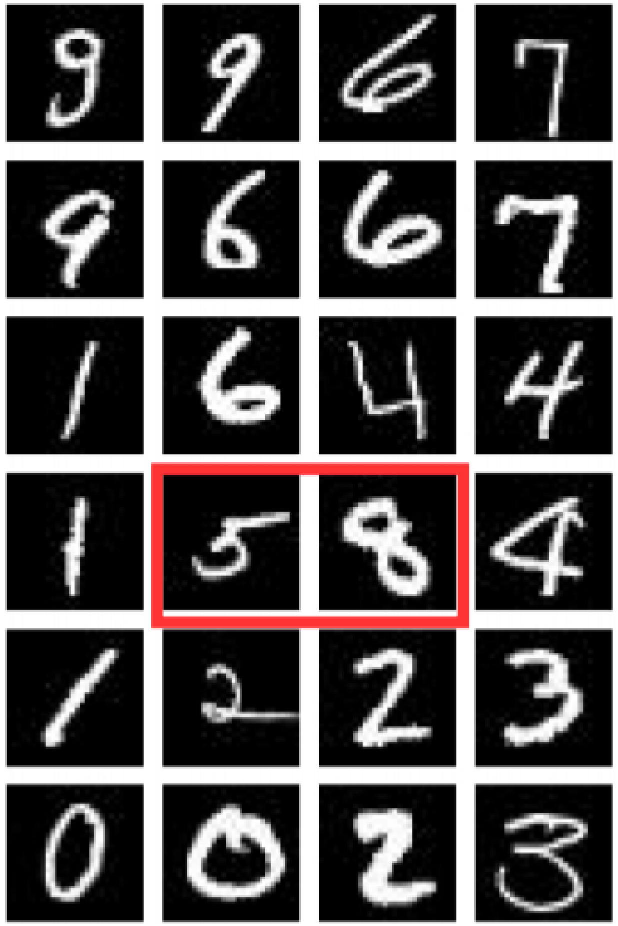
NeurIPS 2021

Active Learning

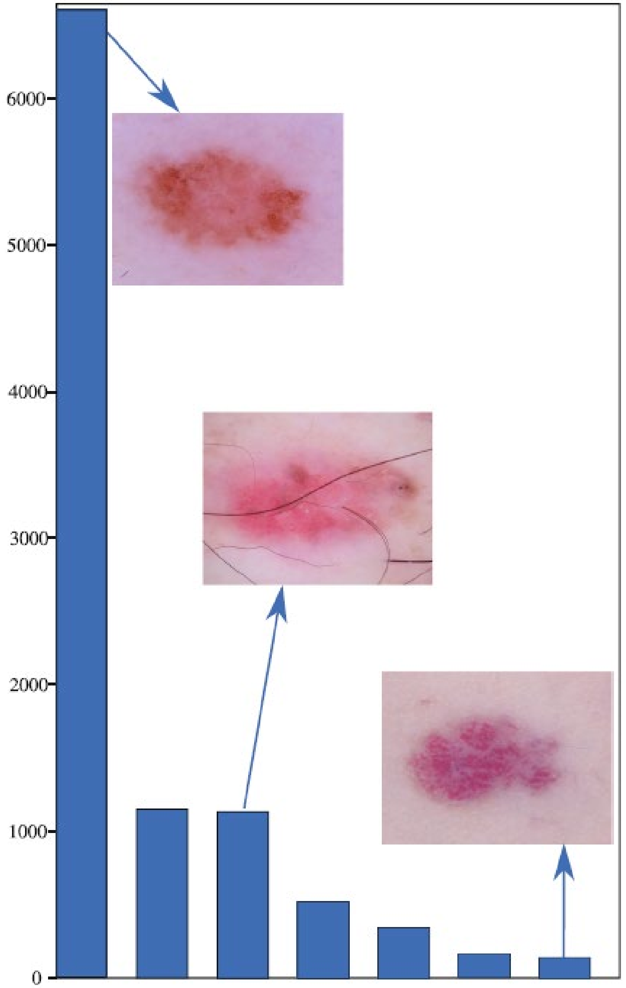


Castillo et al., Industrial statistics and manifold data

Realistic Active Learning Scenarios



(a) Rare Classes

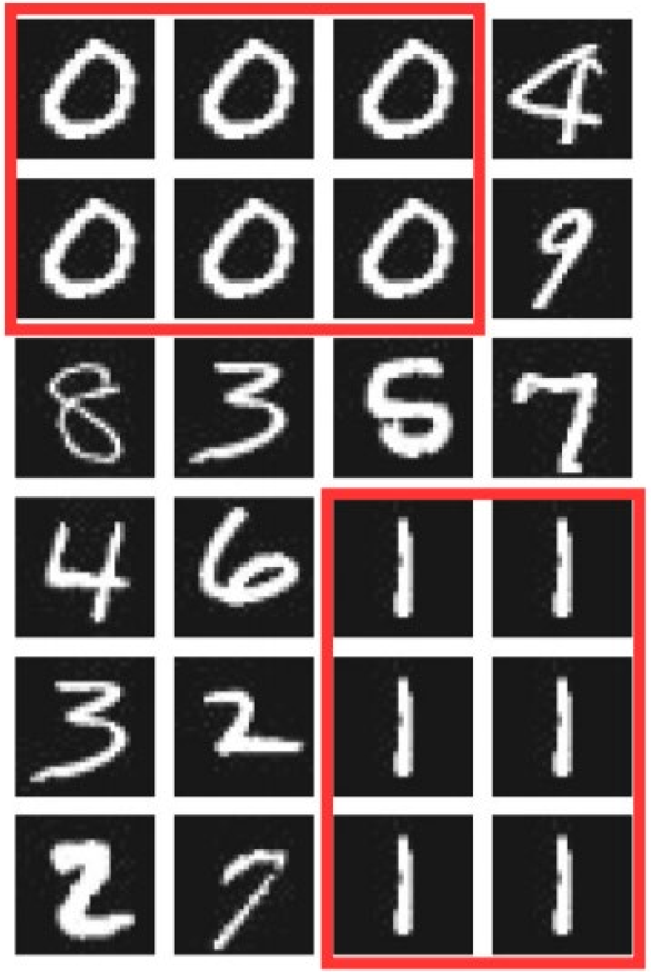


Class Imbalance in ISIC Dataset for Skin Lesions*

Pedestrian in the dark snapshot from a self-driving car**

*Marrakchi et al. MICCAI 2021
**Uber self-driving car crash in Tempe, Arizona.

Realistic Active Learning Scenarios

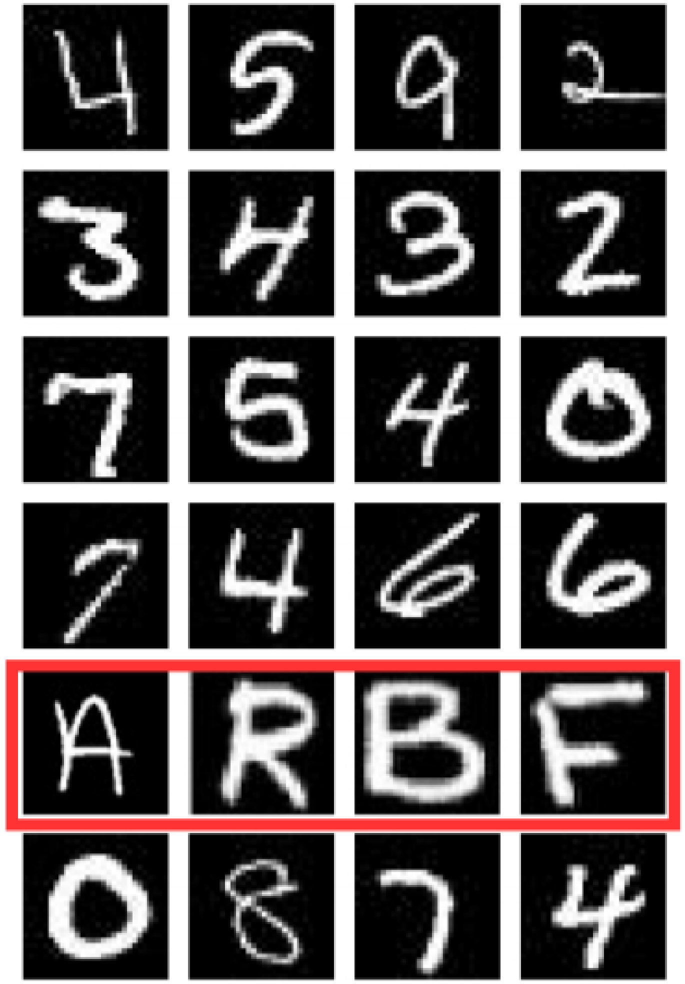


(b) Redundancy

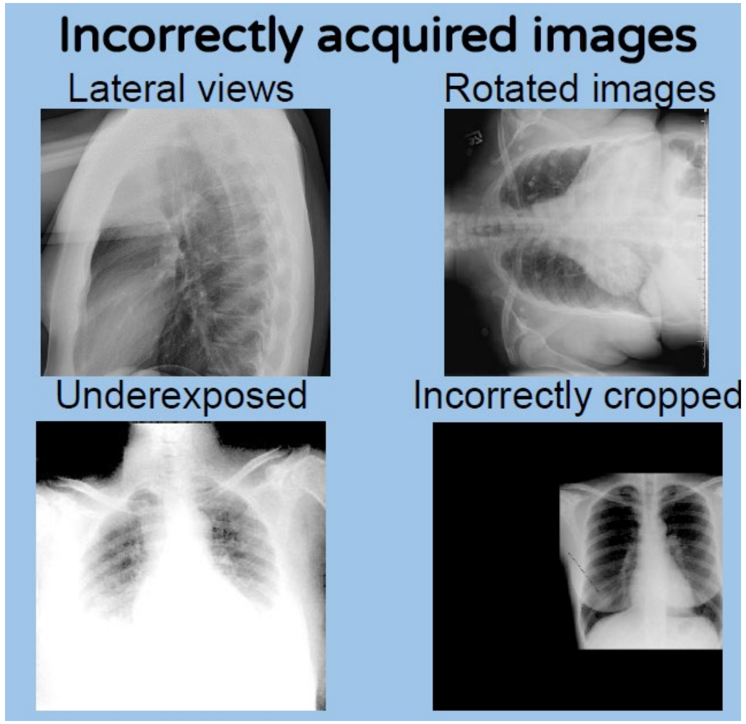
Frames samples from a footage from a self-driving car*

*Source: KITTI

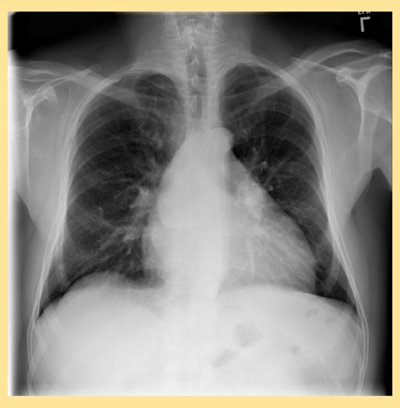
Realistic Active Learning Scenarios



(c) Out-of-distribution



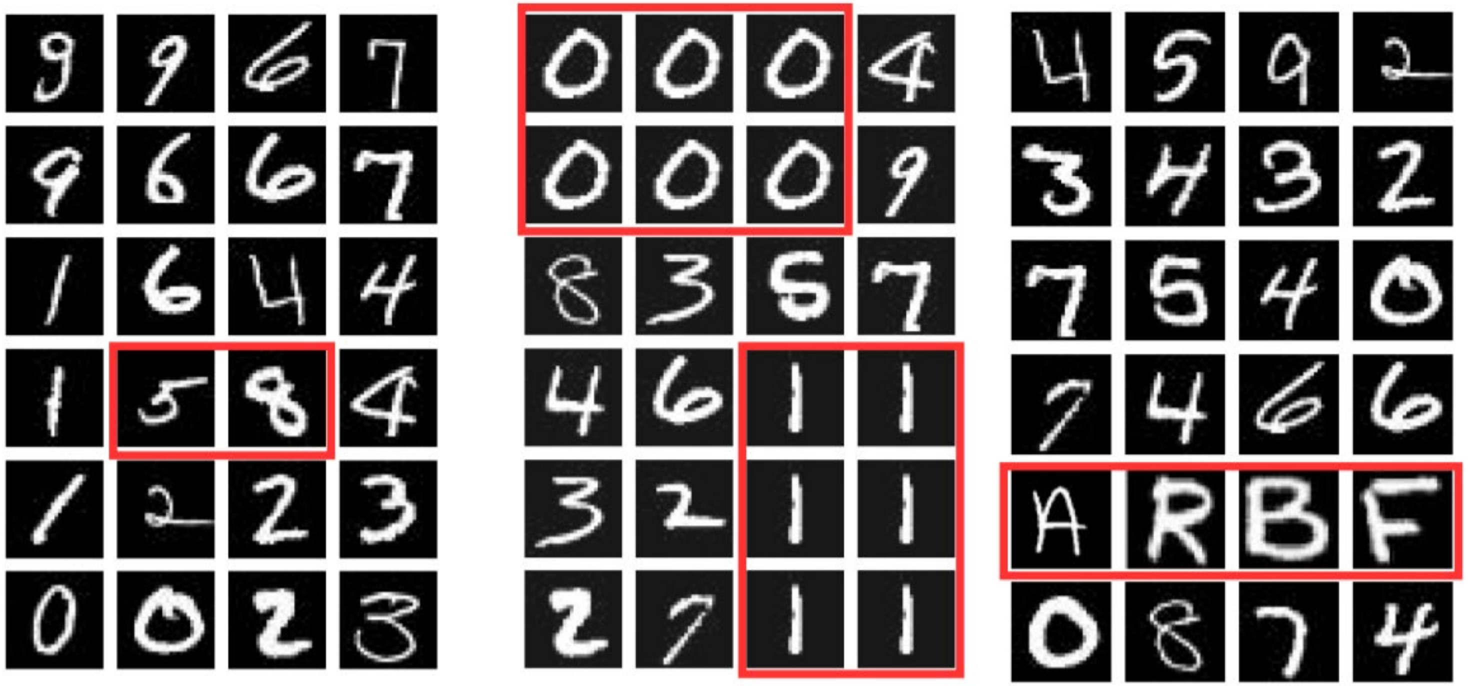
Unfavorable
Out-of-distribution
data points*



Favorable
In-distribution data point*

*Cao et al. A Benchmark of Medical Out of Distribution Detection

The Question

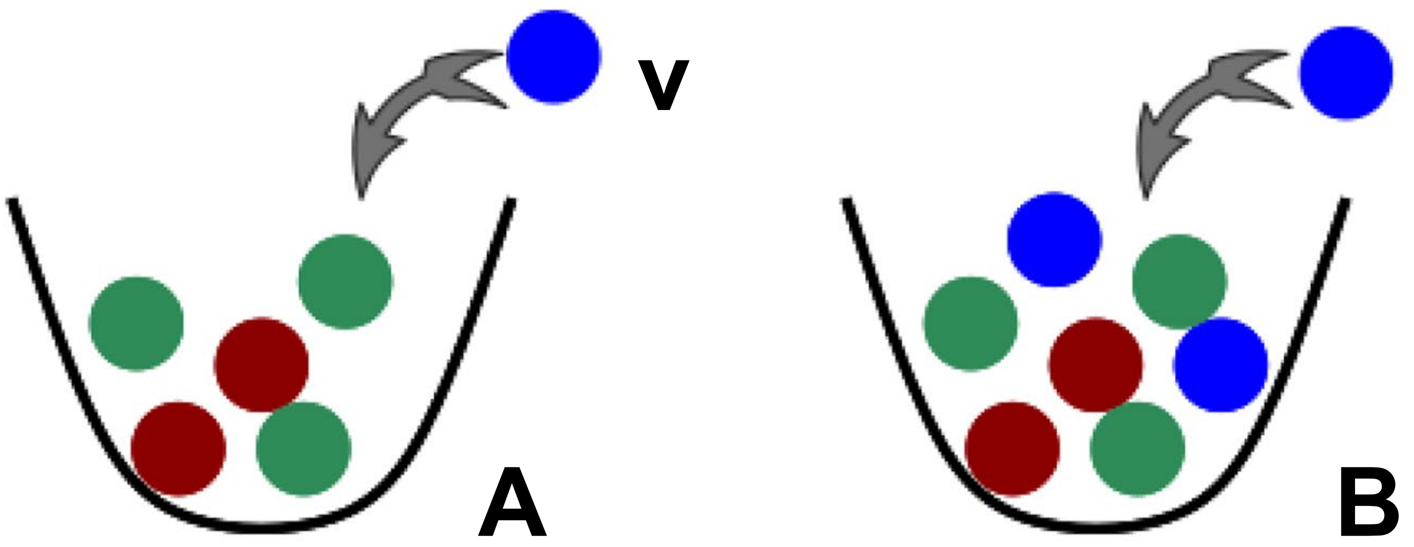


(a) Rare Classes (b) Redundancy (c) Out-of-distribution

Can a machine learning model be trained using a single unified active learning framework that works for a broad spectrum of realistic scenarios?

Submodular Functions

$$f(A \cup v) - f(A) \geq f(B \cup v) - f(B), \text{ if } A \subseteq B$$



$f = \#$ of distinct colors of balls in the urn.

Information Theoretic Concepts

- **Entropy:** Given a set of random variables $X_1 \cdots, X_n$, the Entropy of a **subset** of random variables: $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$. Note that entropy is **submodular.**

Information Theoretic Concepts

- **Entropy:** Given a set of random variables $X_1 \cdots, X_n$, the Entropy of a **subset** of random variables: $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$. Note that entropy is **submodular.**
- **Mutual Information:** Given a set of random variables, $X_1, \cdots, X_n$ and sets $A, B \subseteq V$, the Mutual Information $\mathrm{I}(X_A; X_B) = H(X_A) + H(X_B) - H(X_{A \cup B})$

Information Theoretic Concepts

- **Entropy:** Given a set of random variables $X_1 \cdots, X_n$, the Entropy of a **subset** of random variables: $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$. Note that entropy is **submodular.**
- **Mutual Information:** Given a set of random variables, $X_1, \cdots, X_n$ and sets $A, B \subseteq V$, the Mutual Information $\mathrm{I}(X_A; X_B) = H(X_A) + H(X_B) - H(X_{A \cup B})$
- **Conditional Entropy:** Given a set of random variables, $X_1, \cdots, X_n$ and sets $A, B \subseteq V$, the Conditional Entropy $H(X_A | X_B) = H(X_{A \cup B}) - H(X_B)$

Information Theoretic Concepts

- **Entropy:** Given a set of random variables $X_1 \cdots, X_n$, the Entropy of a **subset** of random variables: $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$. Note that entropy is **submodular.**
- **Mutual Information:** Given a set of random variables, $X_1, \cdots, X_n$ and sets $A, B \subseteq V$, the Mutual Information $\mathrm{I}(X_A; X_B) = H(X_A) + H(X_B) - H(X_{A\cup B})$
- **Conditional Entropy:** Given a set of random variables, $X_1, \cdots, X_n$ and sets $A, B \subseteq V$, the Conditional Entropy $H(X_A|X_B) = H(X_{A\cup B}) - H(X_B)$
- **Conditional Mutual Information:** Given a set of random variables, $X_1, \cdots, X_n$ and sets $A, B, C \subseteq V$, the Conditional Mutual Information $\mathrm{I}(X_A; X_B|X_C) = H(X_A|X_C) + H(X_B|X_C) - H(X_{A\cup B}|X_C)$

Can we replace H with any submodular function?

Can we replace H with any submodular function?

YES!

This gives us the Submodular Information Measures!

Submodular Information Measures (SIM)

- Given a set of data points $V = \{1, \cdots, n\}$, and sets $A, Q \subseteq U$, the **Submodular Mutual Information (SMI)** $I_F(A; Q) = F(A) + F(Q) - F(A \cup Q)$, where the information of a **set** of points is $F(A)$ and F is a submodular function.

Submodular Information Measures (SIM)

- Given a set of data points $V = \{1, \cdots, n\}$**,** and sets $A, Q \subseteq U$**,** the **Submodular Mutual Information (SMI)** $I_F(A; Q) = F(A) + F(Q) - F(A \cup Q)$**,** where the information of a **set** of points is $F(A)$ and F is a submodular function.
- Given a set of data points $V = \{1, \cdots, n\}$, and sets $A, P \subseteq U$, the **Submodular Conditional Gain (SCG)** is $\mathrm{F}(A|P) = F(A \cup P) - F(P)$.

Submodular Information Measures (SIM)

- Given a set of data points $V = \{1, \cdots, n\}$**,** and sets $A, Q \subseteq U$**,** the **Submodular Mutual Information (SMI)** $I_F(A; Q) = F(A) + F(Q) - F(A \cup Q)$**,** where the information of a **set** of points is $F(A)$ and F is a submodular function.
- Given a set of data points $V = \{1, \cdots, n\}$, and sets $A, P \subseteq U$, the **Submodular Conditional Gain (SCG)** is $\mathrm{F}(A|P) = F(A \cup P) - F(P)$.
- Given a set of data points $V = \{1, \cdots, n\}$, and sets $A, Q, P \subseteq U$, the **Submodular Conditional Mutual Information (SCMI)** is $I_F(A; Q|P) = F(A \cup P) + F(Q \cup P) - F(A \cup Q \cup P) - F(P)$.

Submodular Information Measures (SIM)

(a) Instantiations of SMI functions.

SMI	$I_f(\mathcal{A};\mathcal{Q})$
FLVMI	$\sum_{i\in\mathcal{U}} \min(\max_{j\in\mathcal{A}} S_{ij}, \max_{j\in\mathcal{Q}} S_{ij})$
FLQMI	$\sum_{i\in\mathcal{Q}} \max_{j\in\mathcal{A}} S_{ij} + \sum_{i\in\mathcal{A}} \max_{j\in\mathcal{Q}} S_{ij}$
GCMI	$2\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{Q}} S_{ij}$
LOGDETMI	$\log\det(S_{\mathcal{A}}) - \log\det(S_{\mathcal{A}} - S_{\mathcal{A},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{A},\mathcal{Q}}^T)$

(b) Instantiations of SCG and SCMI functions.

SCG	$f(\mathcal{A}\|\mathcal{P})$
FLCG	$\sum_{i\in\mathcal{U}} \max(\max_{j\in\mathcal{A}} S_{ij} - \max_{j\in\mathcal{P}} S_{ij}, 0)$
LogDetCG	$\log\det(S_{\mathcal{A}} - S_{\mathcal{A},\mathcal{P}} S_{\mathcal{P}}^{-1} S_{\mathcal{A},\mathcal{P}}^T)$
GCCG	$f(\mathcal{A}) - 2\sum_{i\in\mathcal{A}, j\in\mathcal{P}} S_{ij}$

SCMI	$I_f(\mathcal{A};\mathcal{Q}\|\mathcal{P})$
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	$\log \frac{\det(I - S_{\mathcal{P}}^{-1} S_{\mathcal{P},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{P},\mathcal{Q}}^T)}{\det(I - S_{\mathcal{A}\cup\mathcal{P}}^{-1} S_{\mathcal{A}\cup\mathcal{P},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{A}\cup\mathcal{P},\mathcal{Q}}^T)}$

Iyer at al., ALT 2021

Submodular Mutual Information (SMI)

SMI	$I_f(\mathcal{A}; \mathcal{Q})$
FLVMI	$\sum_{i \in \mathcal{U}} \min(\max_{j \in \mathcal{A}} S_{ij}, \max_{j \in \mathcal{Q}} S_{ij})$
FLQMI	$\sum_{i \in \mathcal{Q}} \max_{j \in \mathcal{A}} S_{ij} + \sum_{i \in \mathcal{A}} \max_{j \in \mathcal{Q}} S_{ij}$
GCMI	$2 \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{Q}} S_{ij}$
LOGDETMI	$\log\det(S_{\mathcal{A}}) - \log\det(S_{\mathcal{A}} - S_{\mathcal{A},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{A},\mathcal{Q}}^T)$

Submodular Conditional Gain (SCG)

SCG	$f(\mathcal{A}\|\mathcal{P})$
FLCG	$\sum_{i \in \mathcal{U}} \max(\max_{j \in \mathcal{A}} S_{ij} - \max_{j \in \mathcal{P}} S_{ij}, 0)$
LogDetCG	$\log\det(S_{\mathcal{A}} - S_{\mathcal{A},\mathcal{P}} S_{\mathcal{P}}^{-1} S_{\mathcal{A},\mathcal{P}}^{T})$
GCCG	$f(\mathcal{A}) - 2 \sum_{i \in \mathcal{A}, j \in \mathcal{P}} S_{ij}$

Submodular Conditional Mutual Information (SCMI)

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	$\log \frac{\det(I - S_{\mathcal{P}}^{-1} S_{\mathcal{P},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{P},\mathcal{Q}}^{T})}{\det(I - S_{\mathcal{A}\cup\mathcal{P}}^{-1} S_{\mathcal{A}\cup\mathcal{P},Q} S_{\mathcal{Q}}^{-1} S_{\mathcal{A}\cup\mathcal{P},Q}^{T})}$

Relationship between SIM

$$\max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} I_f(\mathcal{A}; \mathcal{Q} | \mathcal{P})$$

Function	Setting	Realistic Scenario
Submodular	$\mathcal{Q} \leftarrow \mathcal{U}, \mathcal{P} \leftarrow \emptyset$	Standard AL
SMI	$\mathcal{Q} \leftarrow \mathcal{Q}, \mathcal{P} \leftarrow \emptyset$	Imbalance, OOD
SCG	$\mathcal{Q} \leftarrow \emptyset, \mathcal{P} \leftarrow \mathcal{P}$	Redundancy
SCMI	$\mathcal{Q} \leftarrow \mathcal{Q}, \mathcal{P} \leftarrow \mathcal{P}$	OOD

SIMILAR: Unified AL Framework

Require: Initial labeled set of data points: $\mathcal{L}$, large unlabeled dataset: $\mathcal{U}$, loss function $\mathcal{H}$ for learning model $\mathcal{M}$, batch size: B, number of selection rounds: N

1. **for** selection round $i = 1 : N$ **do**
2. Train model $\mathcal{M}$ with loss $\mathcal{H}$ on the current labeled set $\mathcal{L}$ and obtain parameters θ_i
3. Using model parameters θ_i, compute gradients using hypothesized labels $\{\nabla_{\theta_i}\mathcal{H}(x_j, \hat{y}_j, \theta_i), \forall j \in \mathcal{U}\}$ and obtain a similarity matrix X
4. Instantiate a submodular function f based on X
5. $\mathcal{A}_i \leftarrow \mathrm{argmax}_{\mathcal{A}\subseteq\mathcal{U}, |\mathcal{A}|\leq B}\ I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P})$ (Optimize SCMI with an appropriate choice of $\mathcal{Q}$ and $\mathcal{P}$, see Tab. 1)
6. Get labels $L(\mathcal{A}_i)$ for batch $\mathcal{A}_i$ and $\mathcal{L} \leftarrow \mathcal{L} \cup L(\mathcal{A}_i), \mathcal{U} \leftarrow \mathcal{U} - \mathcal{A}_i$
7. **end for**
8. **return** trained model $\mathcal{M}$ and parameters θ_N

SIMILAR: Unified AL Framework

Require: Initial labeled set of data points: $\mathcal{L}$, large unlabeled dataset: $\mathcal{U}$, loss function $\mathcal{H}$ for learning model $\mathcal{M}$, batch size: B, number of selection rounds: N

1. **for** selection round $i = 1 : N$ **do**
2. Train model $\mathcal{M}$ with loss $\mathcal{H}$ on the current labeled set $\mathcal{L}$ and obtain parameters θ_i
3. Using model parameters θ_i, compute gradients using hypothesized labels $\{\nabla_{\theta_i}\mathcal{H}(x_j, \hat{y}_j, \theta_i), \forall j \in \mathcal{U}\}$ and obtain a similarity matrix X
4. Instantiate a submodular function f based on X
5. $\mathcal{A}_i \leftarrow \text{argmax}_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P})$ (Optimize SCMI with an appropriate choice of $\mathcal{Q}$ and $\mathcal{P}$, see Tab. 1)
6. Get labels $L(\mathcal{A}_i)$ for batch $\mathcal{A}_i$ and $\mathcal{L} \leftarrow \mathcal{L} \cup L(\mathcal{A}_i), \mathcal{U} \leftarrow \mathcal{U} - \mathcal{A}_i$
7. **end for**
8. **return** trained model $\mathcal{M}$ and parameters θ_N

SIMILAR: Unified AL Framework

Require: Initial labeled set of data points: $\mathcal{L}$, large unlabeled dataset: $\mathcal{U}$, loss function $\mathcal{H}$ for learning model $\mathcal{M}$, batch size: B, number of selection rounds: N

1. **for** selection round $i = 1 : N$ **do**
2. Train model $\mathcal{M}$ with loss $\mathcal{H}$ on the current labeled set $\mathcal{L}$ and obtain parameters θ_i
3. Using model parameters θ_i, compute gradients using hypothesized labels $\{\nabla_{\theta_i}\mathcal{H}(x_j, \hat{y}_j, \theta_i), \forall j \in \mathcal{U}\}$ and obtain a similarity matrix X
4. Instantiate a submodular function f based on X
5. $\mathcal{A}_i \leftarrow \mathrm{argmax}_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B}\, I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P})$ (Optimize SCMI with an appropriate choice of $\mathcal{Q}$ and $\mathcal{P}$, see Tab. 1)
6. Get labels $L(\mathcal{A}_i)$ for batch $\mathcal{A}_i$ and $\mathcal{L} \leftarrow \mathcal{L} \cup L(\mathcal{A}_i), \mathcal{U} \leftarrow \mathcal{U} - \mathcal{A}_i$
7. **end for**
8. **return** trained model $\mathcal{M}$ and parameters θ_N

SIMILAR: Unified AL Framework

Require: Initial labeled set of data points: $\mathcal{L}$, large unlabeled dataset: $\mathcal{U}$, loss function $\mathcal{H}$ for learning model $\mathcal{M}$, batch size: B, number of selection rounds: N

for selection round $i = 1 : N$ **do**

    Train model $\mathcal{M}$ with loss $\mathcal{H}$ on the current labeled set $\mathcal{L}$ and obtain parameters θ_i

    Using model parameters θ_i, compute gradients using hypothesized labels $\{\nabla_{\theta_i}\mathcal{H}(x_j, \hat{y}_j, \theta_i), \forall j \in \mathcal{U}\}$ and obtain a similarity matrix X

    Instantiate a submodular function f based on X

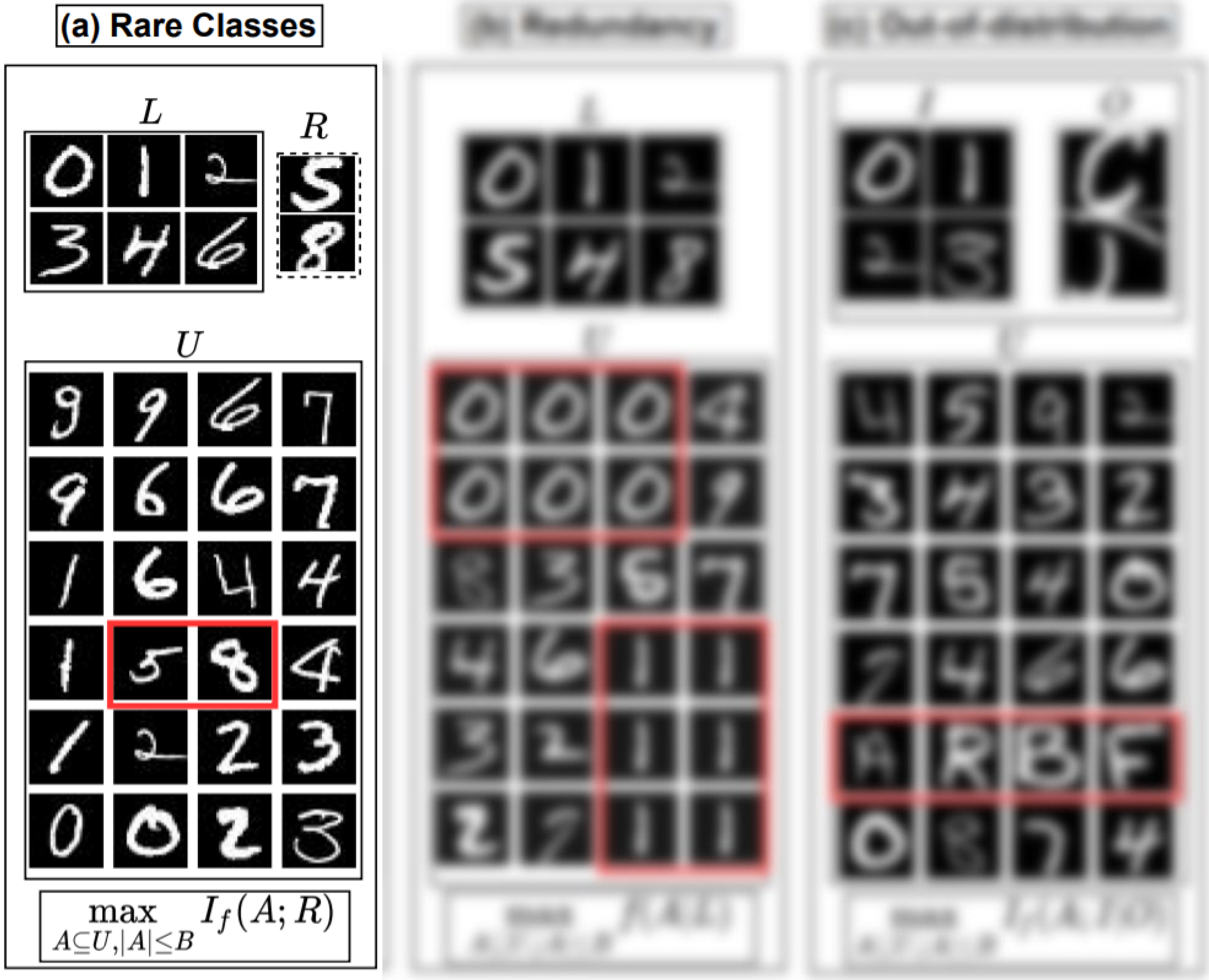
    $\mathcal{A}_i \leftarrow \text{argmax}_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B}\, I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P})$ (Optimize SCMI with an appropriate choice of $\mathcal{Q}$ and $\mathcal{P}$, see Tab. 1)

    Get labels $L(\mathcal{A}_i)$ for batch $\mathcal{A}_i$ and $\mathcal{L} \leftarrow \mathcal{L} \cup L(\mathcal{A}_i), \mathcal{U} \leftarrow \mathcal{U} - \mathcal{A}_i$

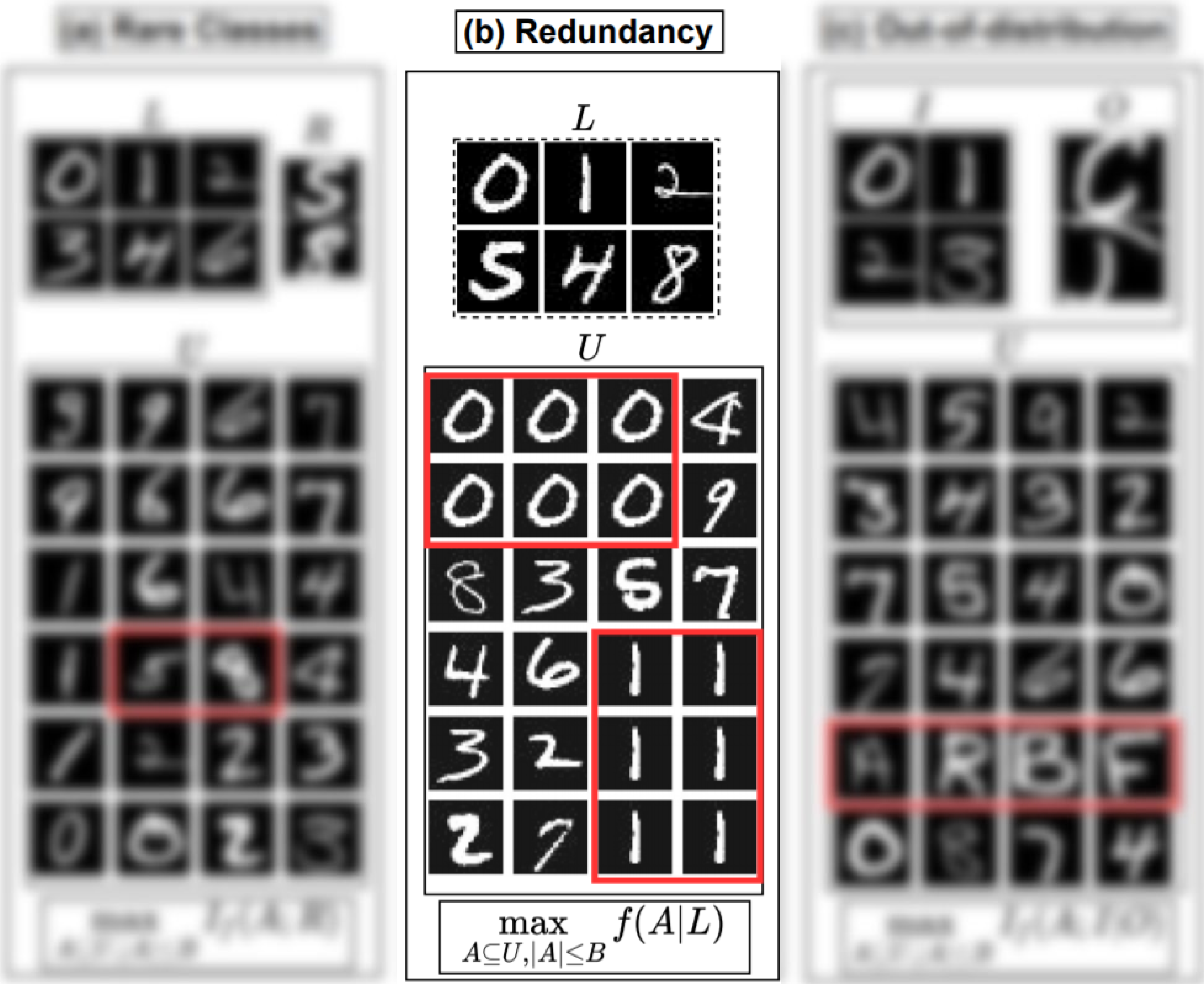
end for

return trained model $\mathcal{M}$ and parameters θ_N

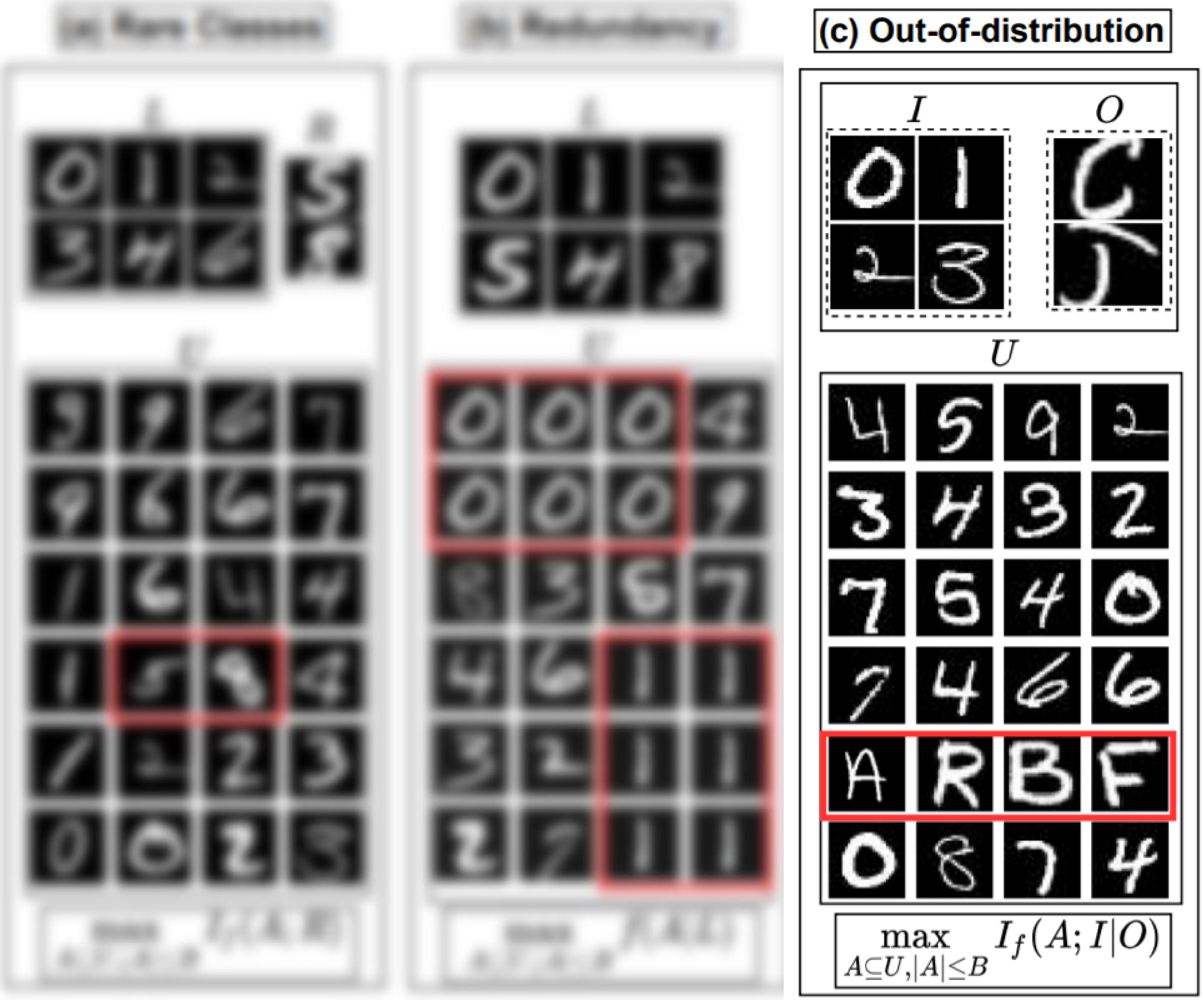
Choices of Query and Conditioning Sets



Choices of Query and Conditioning Sets



Choices of Query and Conditioning Sets



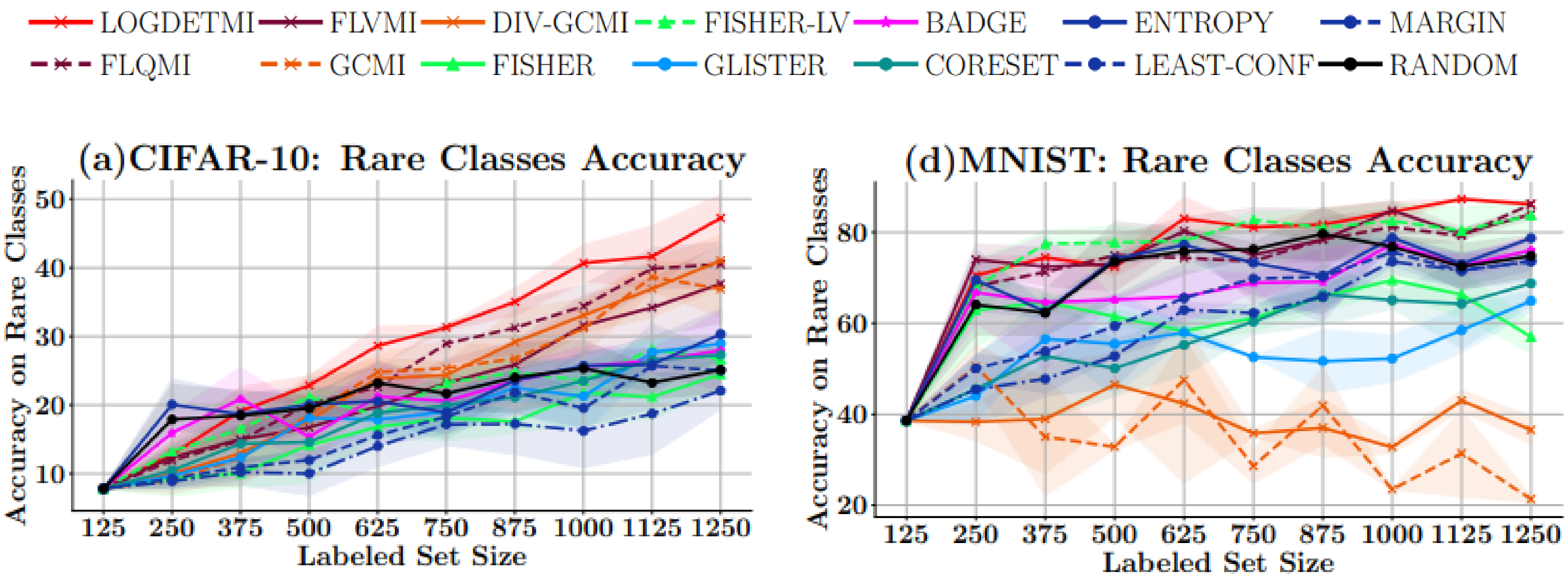
Choices of Query and Conditioning Sets for Multiple Co-occurring Realistic Scenarios

$$\max_{\mathcal{A}\subseteq\mathcal{U},|\mathcal{A}|\leq B} I_f(\mathcal{A};\mathcal{Q}|\mathcal{P})$$

Function	Setting	Realistic Scenario
$I_f(\mathcal{A}:\mathcal{R}\vert\mathcal{O})$	$\mathcal{Q}\leftarrow\mathcal{R},\mathcal{P}\leftarrow\mathcal{O}$	Rare classes + OOD
$I_f(\mathcal{A};\mathcal{R}\vert\mathcal{L}-\tilde{\mathcal{R}})$	$\mathcal{Q}\leftarrow\mathcal{R},\mathcal{P}\leftarrow\mathcal{L}-\tilde{\mathcal{R}}$	Rare classes + Redundancy
$I_f(\mathcal{A};\mathcal{I}\vert\mathcal{O}\cup\mathcal{I}^*)$	$\mathcal{Q}\leftarrow\mathcal{I},\mathcal{P}\leftarrow\mathcal{O}\cup\mathcal{I}^*$	Redundancy + OOD

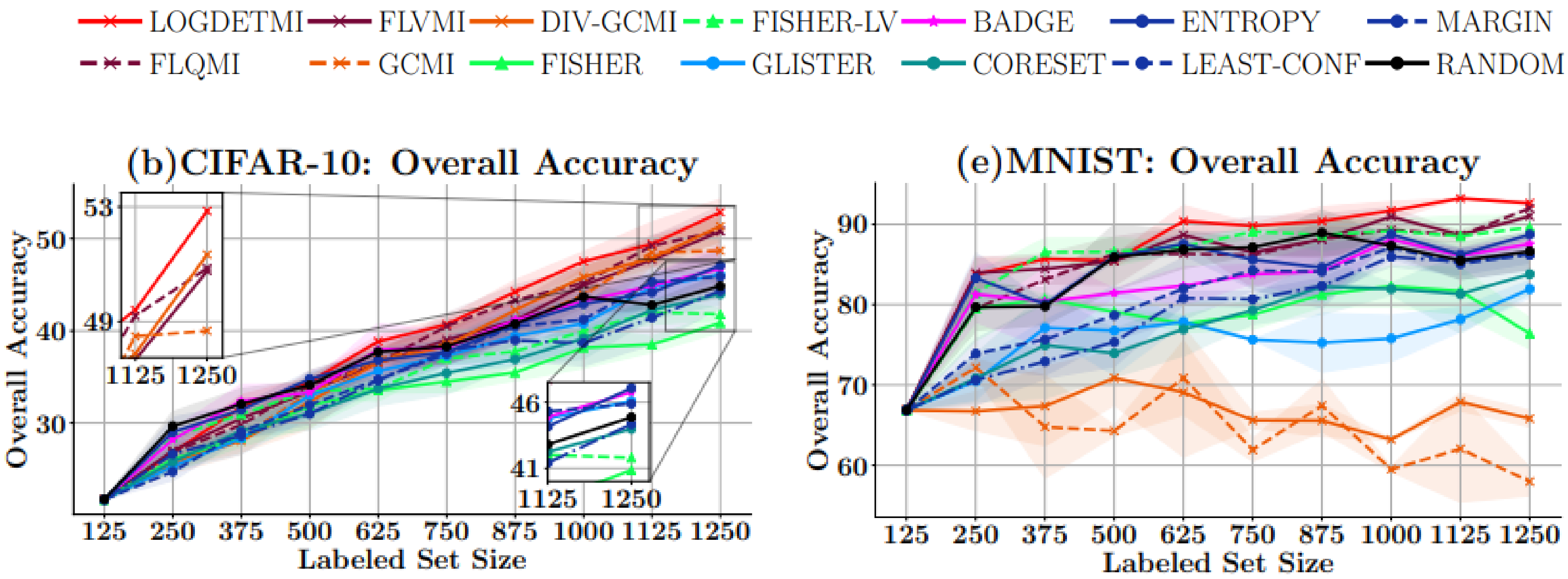
- For Rare classes + Redundancy: $\tilde{R}$ is the subset of data points from the labeled set L that belong to the rare classes.
- For Redundancy + OOD: Kernel for I^* is computed using an exponential kernel to penalize similar samples in I.

Results: AL with Rare Classes



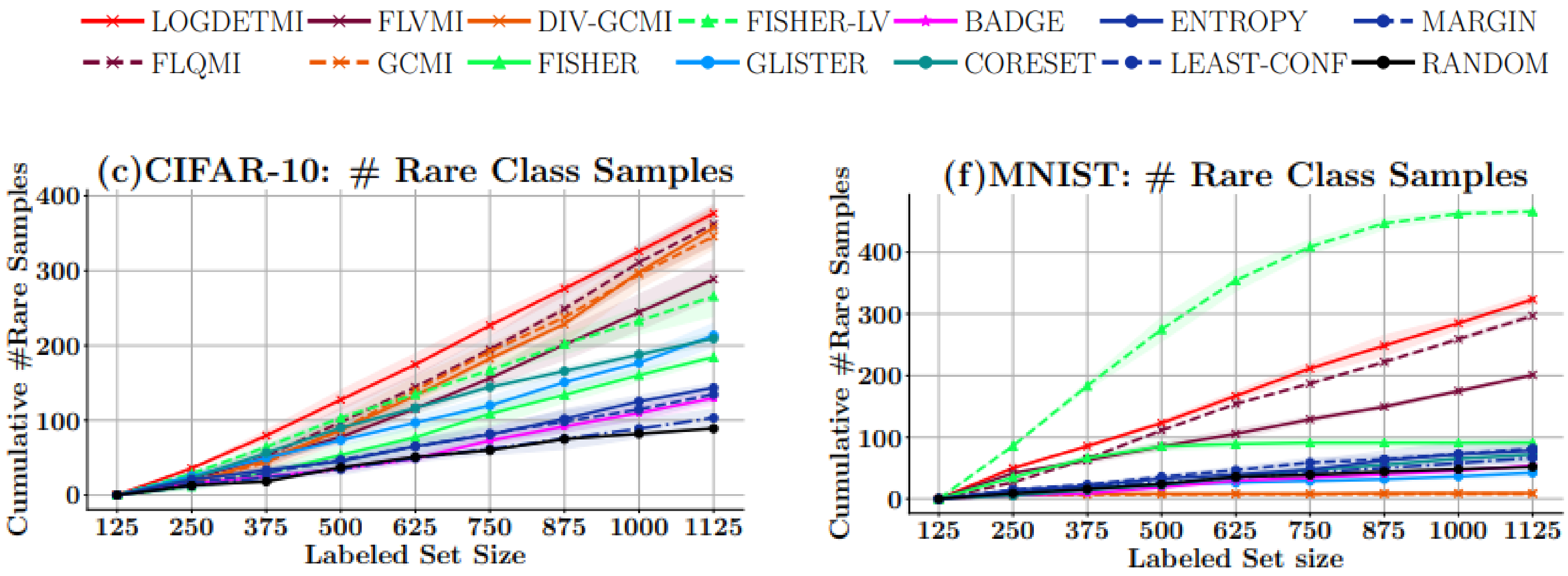
- SMI based functions not only consistently outperforms all baselines by ~ 10 − 18% in terms of average accuracy on rare classes.
- FLQMI and LOGDETMI which balance between diversity and relevance perform better than GCMI which only models relevance.

Results: AL with Rare Classes



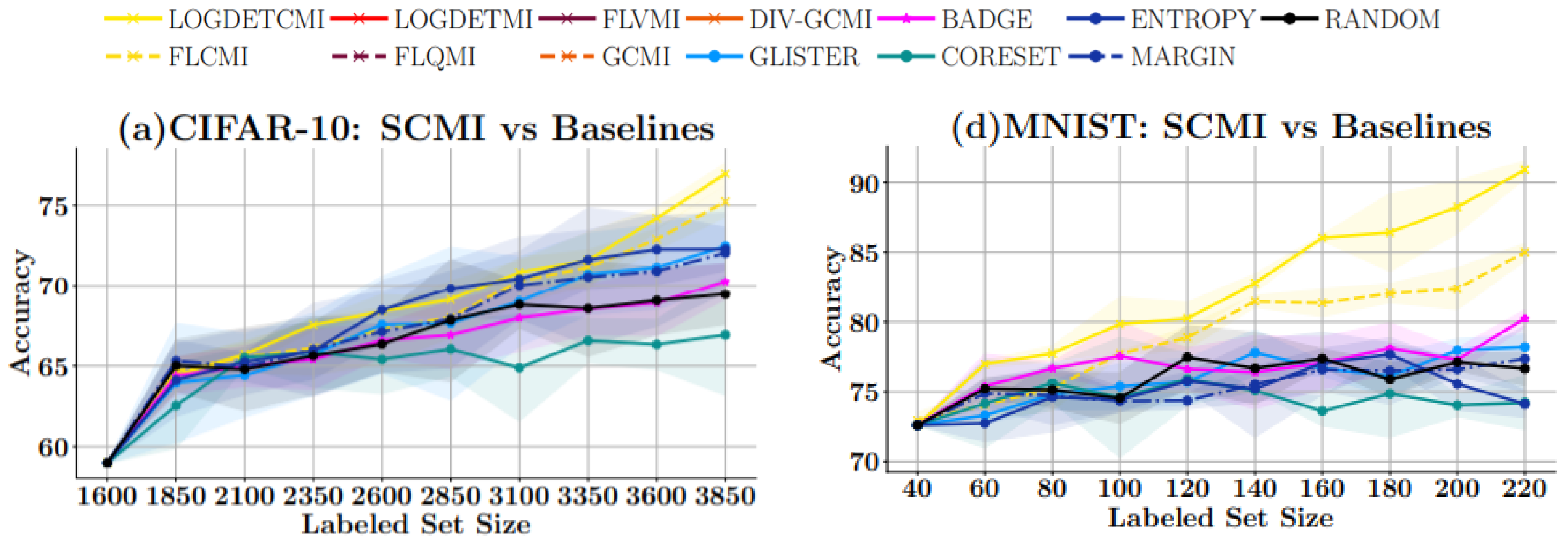
SMI based functions not only consistently outperforms all baselines by by ~ 5 − 10% in terms of overall accuracy.

Results: AL with Rare Classes



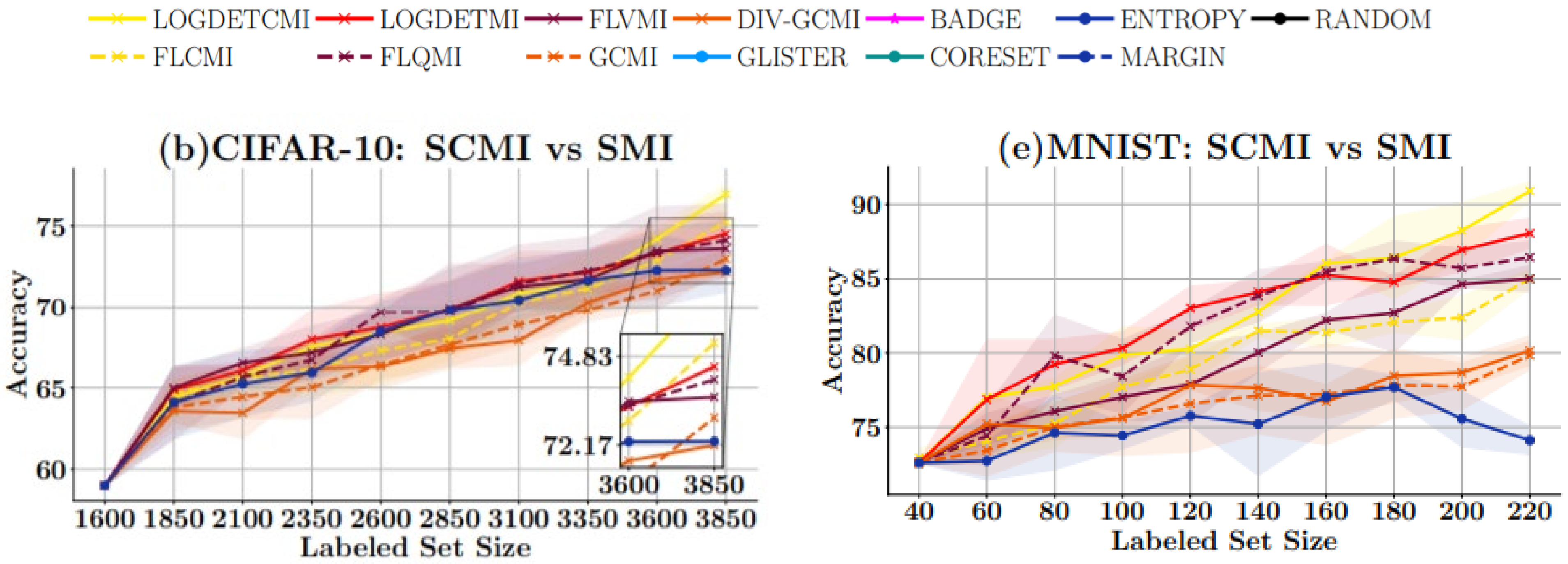
The gain in performance is because SMI functions pick the greatest number of diverse datapoints from the rare classes.

Results: AL with OOD Data



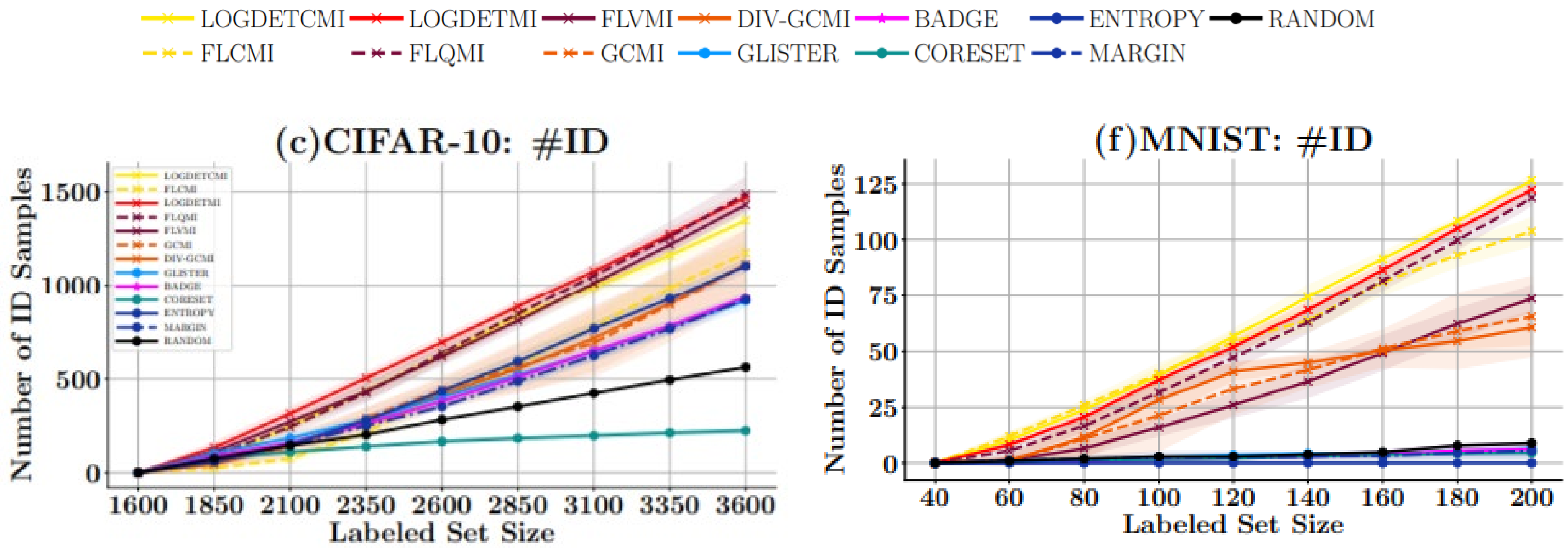
- SCMI based acquisition functions significantly outperform existing AL approaches by ~ 5 − 10%
- Existing acquisition functions have a high variance which is undesirable in real-world deployment scenarios. Our SCMI based acquisition functions show the lowest variance in training.

Results: AL with OOD Data



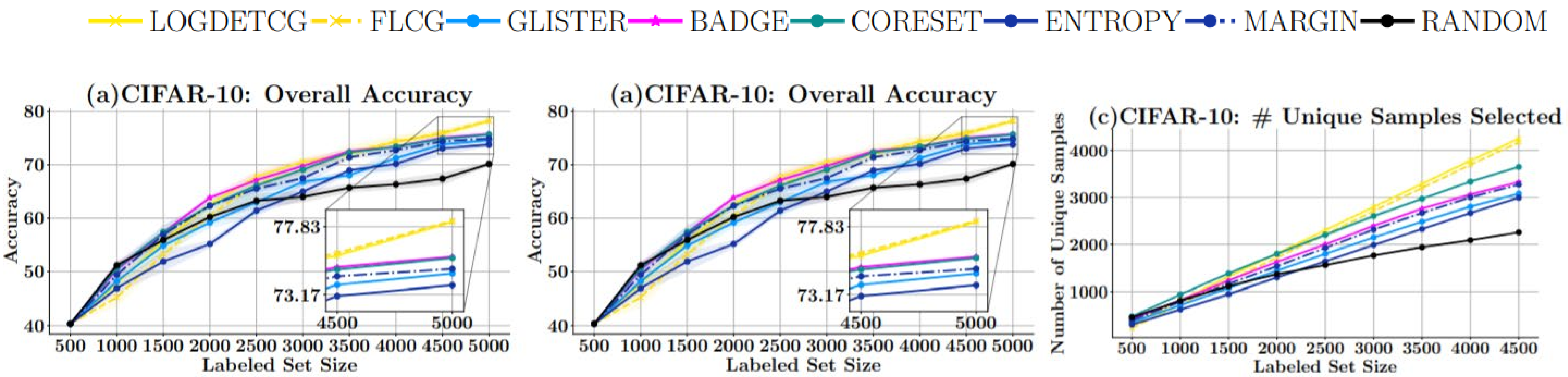
SCMI functions show ~2 − 3% improvement over SMI as the conditioning becomes stronger. This is because SCMI tend to select more in-distribution points compared to SMI.

Results: AL with OOD Data



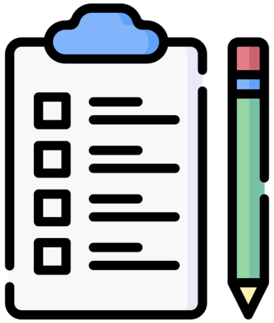
SMI and SCMI tend to select more in-distribution points compared to baselines.

Results: AL with Redundancy



- As expected, the diversity and uncertainty based methods outperform random.
- We observe that the SCG functions significantly outperform all baselines by ~ 3 − 5% in the later active learning rounds as the conditioning gets stronger.
- We observe this gain because SCG based acquisition functions select significantly more unique points than other methods.

Conclusion



- We proposed a unified active learning framework SIMILAR using the submodular information functions.
- We showed the applicability of the framework in three realistic scenarios for active learning, namely rare classes, redundancy, and out of distribution data.
- In each case, we observed that the functions in SIMILAR significantly outperform existing baselines in each of these tasks.
- Our real-world experiments on MNIST, CIFAR-10, and ImageNet show that many of the SIM functions (specifically the LOGDET and FL variants) yield $\sim 5\% - 18\%$ gain compared to existing baselines particularly in the rare class scenario and $\sim 5\% - 10\%$ OOD scenarios.

Thank You

For more details, do visit our poster.