SIMILAR: Submodular Information Measures Based Active Learning In Realistic Scenarios

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, Rishabh Iyer

THE UNIVERSITY OF TEXAS AT DALLAS

NEURAL INFORMATION PROCESSING SYSTEMS

INTRODUCTION

- Active learning (AL) has proven to be useful for minimizing labeling costs by selecting the most informative samples.
- However, existing AL methods do not work well in realistic scenarios such as imbalance or rare classes, out-of-distribution (OOD) data in the unlabeled set, and redundancy.
- We propose *SIMILAR* using Submodular Mutual Information Measures (SIM). *SIMILAR* acts as a one-stop solution for AL.
- We show that *SIMILAR* significantly outperforms existing AL algorithms by as much as ≈5%–18% in the case of rare classes and ≈5%–10% in the case of OOD data on multiple image classification datasets.

PROBLEM FORMULATION

Goal: A unified active learning framework that works for a broad spectrum of realistic scenarios.

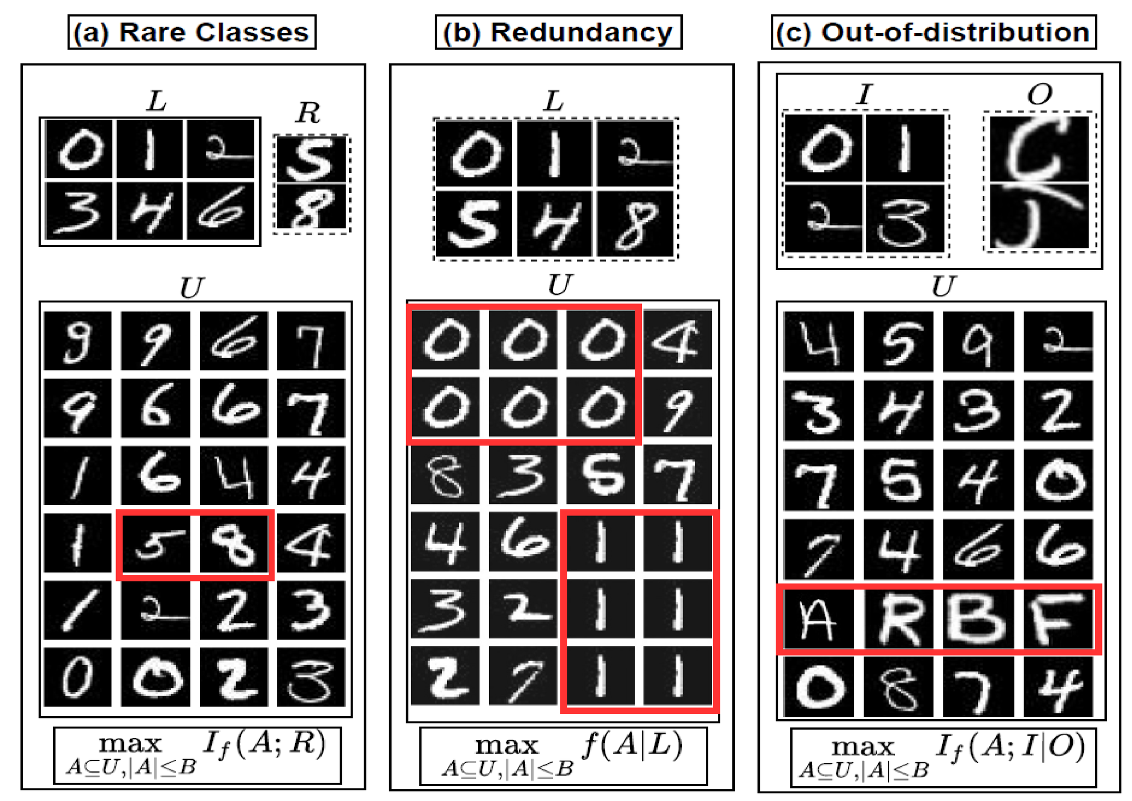
- The Submodular Conditional Mutual Information (SCMI) is defined as $I_f(A;Q|P) = f(A \cup P) + f(Q \cup P) - f(A \cup Q \cup P) - f(P)$. It jointly models the similarity between A and Q, and their dissimilarity with P.
- Depending on the realistic scenario, we instantiate SCMI with appropriate choices of Q and P as follows:

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	$\mathcal{Q} \leftarrow \mathcal{U}, \mathcal{P} \leftarrow \mathcal{P}$	Redundancy
SCMI	$\mathcal{Q} \leftarrow \mathcal{Q}, \mathcal{P} \leftarrow \mathcal{P}$	OOD

- To find an optimal subset we optimize:

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

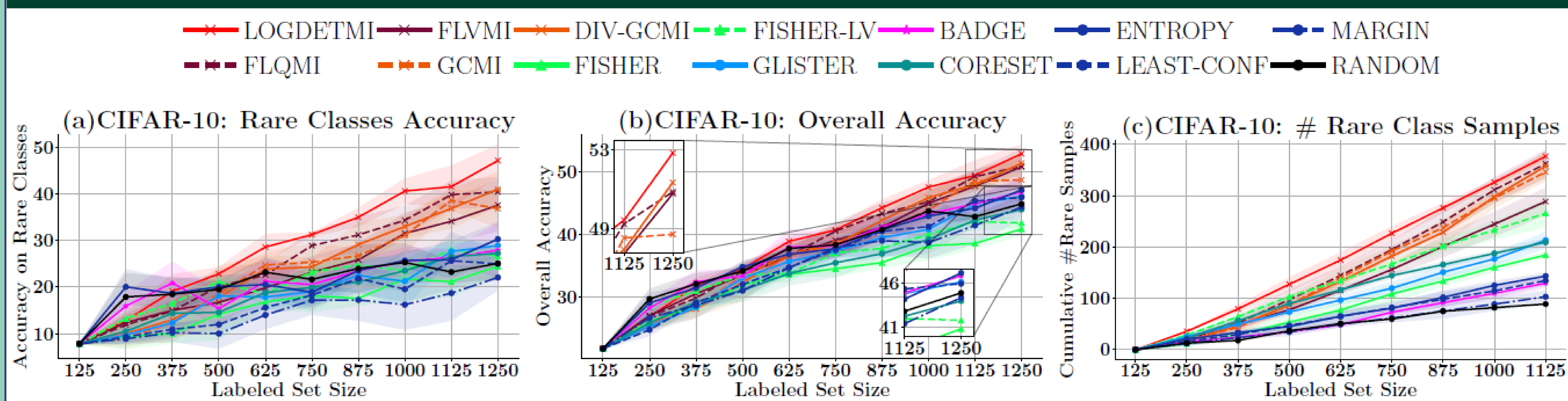
Application of *SIMILAR* to realistic scenarios

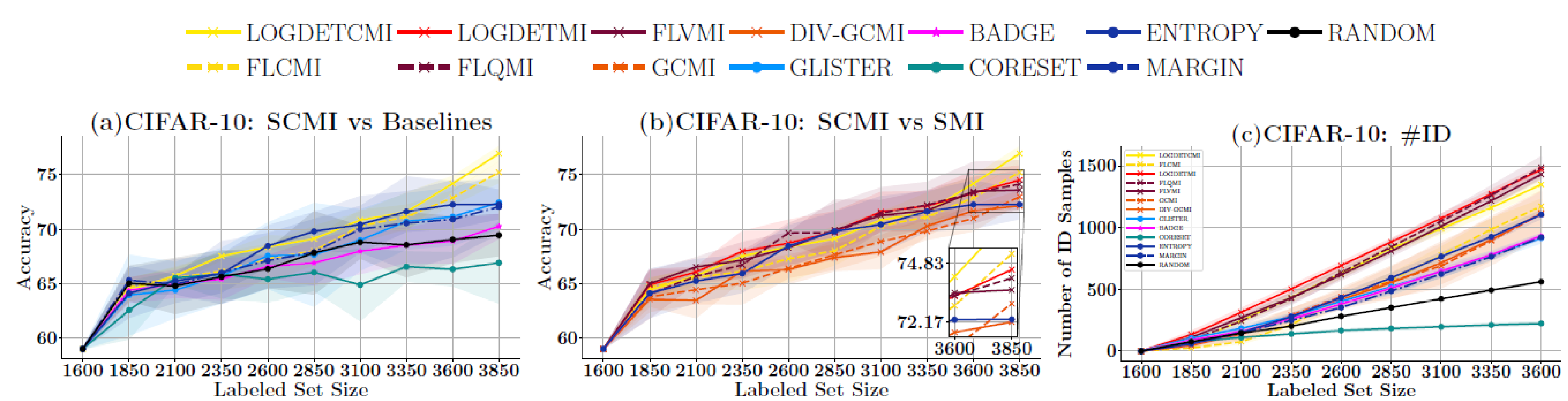


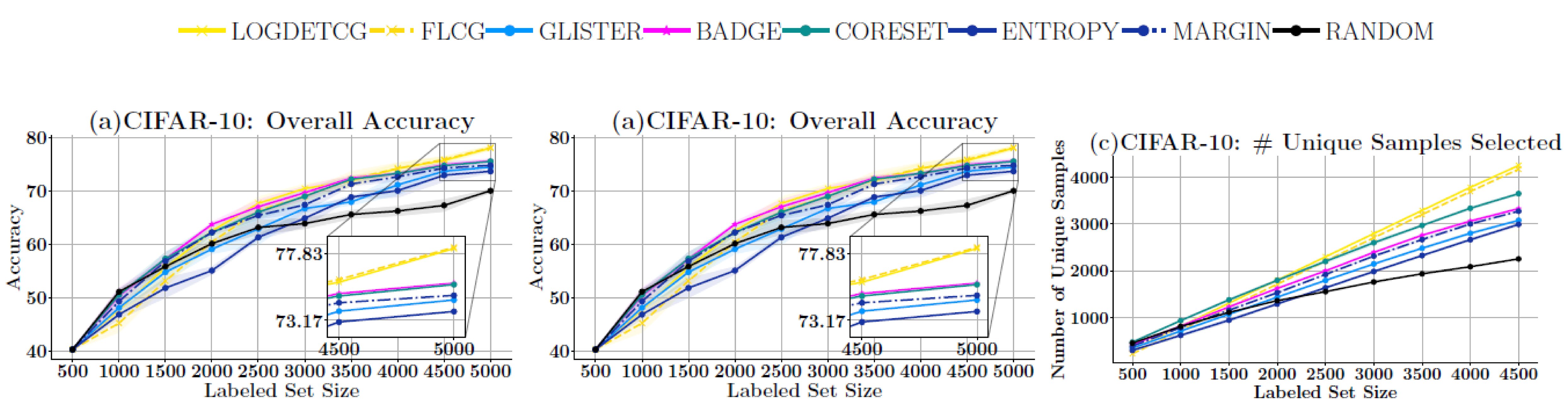
Motivating scenarios for realistic active learning and illustration of appropriate choices of Q and P

- SIMILAR finds rare digits 5,8 ∈ U by optimizing the SMI function $I_f(A;R)$ with R containing 5, 8 as *queries*.
- *SIMILAR* selects samples from U which are diverse among themselves and also diverse w.r.t those in L by optimizing $f(A|L)$ (here, we want to avoid digits 0,1 ∈ U because they are present in L)
- *SIMILAR selects digits (in-distribution) and avoid alphabets (out-of-distribution) in U by optimizing $I_f(A;I|O)$, where I are ID labeled points and O are OOD points selected so far.*

RESULTS







SIMILAR significantly outperforms existing active learning algorithms by as much as 5%-18% in the case of rare classes and 5%-10% in the case of out-of-distribution data on CIFAR-10 image classification.

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
3: Using model parameters θ_i, compute gradients using hypothesized labels $\{\nabla_\theta \mathcal{H}(x_j, \hat{y}_j, \theta), \forall j \in \mathcal{U}\}$ 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.

CONCLUSIONS

- We demonstrate the effectiveness of *SIMILAR* in three realistic scenarios for active learning, namely rare classes, redundancy, and out of distribution data.
- Our real-world experiments show that many of the SIM functions (specifically the LOGDET and FL variants) yield 5%-18% gain compared to existing baselines particularly in the rare class scenario and 5%-10% OOD scenarios.

PAPER

SCAN ME

Get the paper for more technical details and results: https://arxiv.org/pdf/2107.00717.pdf