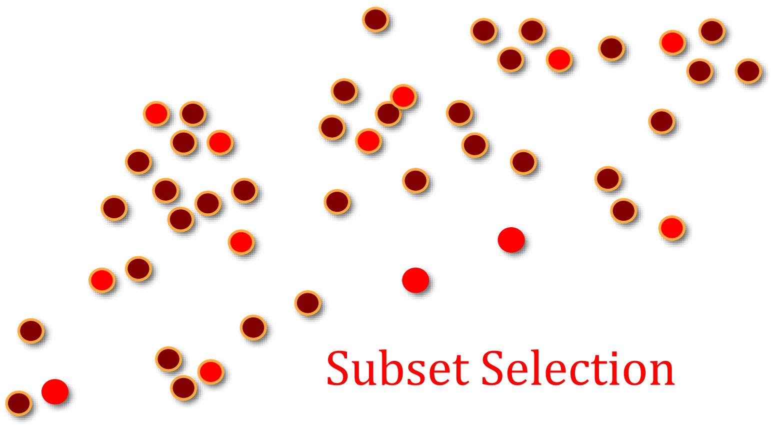
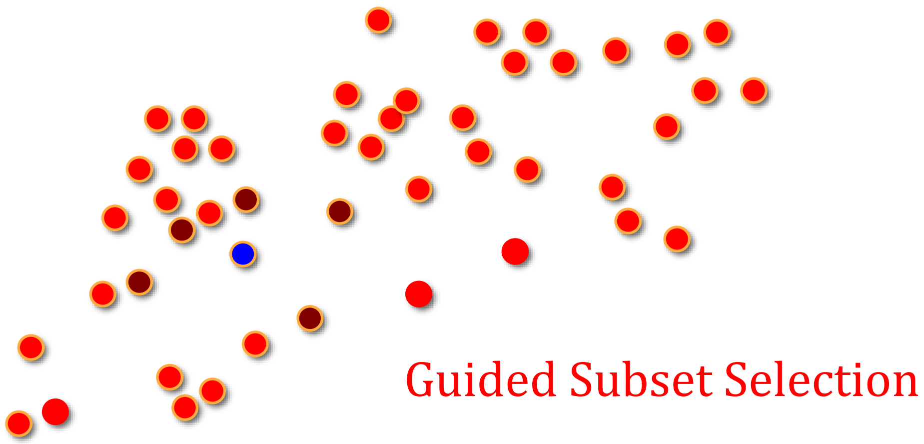
PRISM: A Rich Class of Parameterized Submodular Information Measures for Guided Subset Selection

Suraj Kothawade*, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, Rishabh Iyer

*suraj.kothawade@utdallas.edu

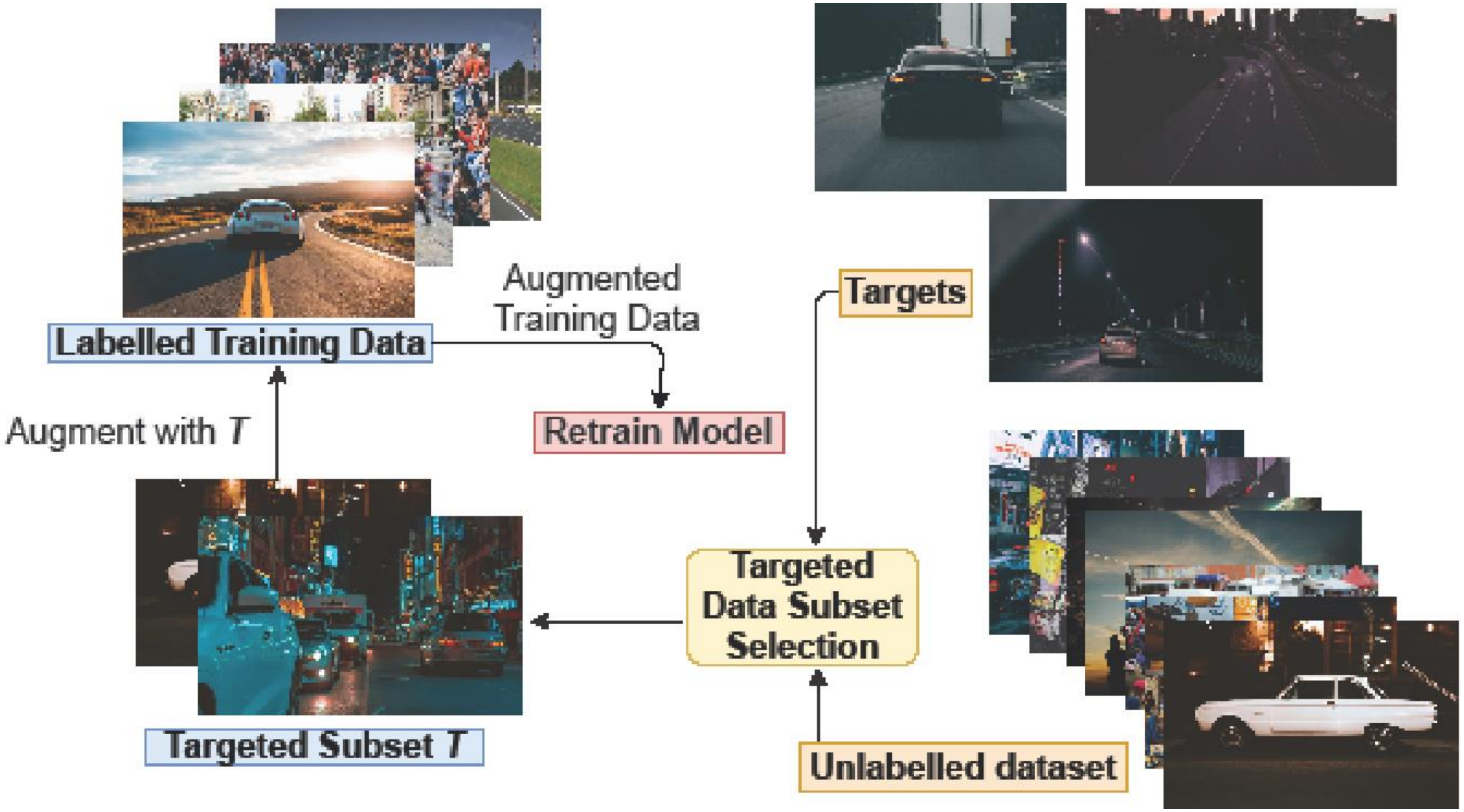
Guided Subset Selection



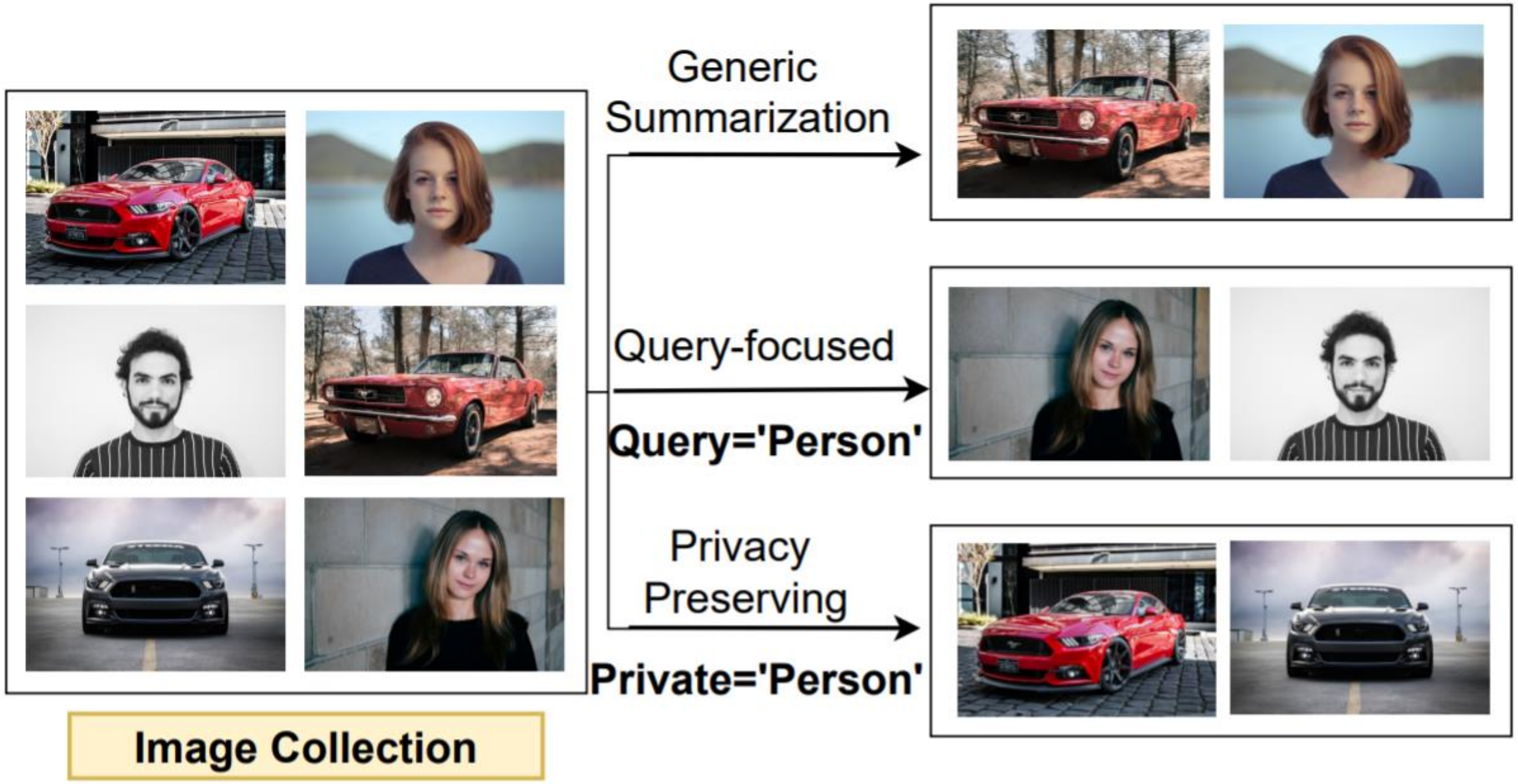


Examples of Targets

Targeted Learning

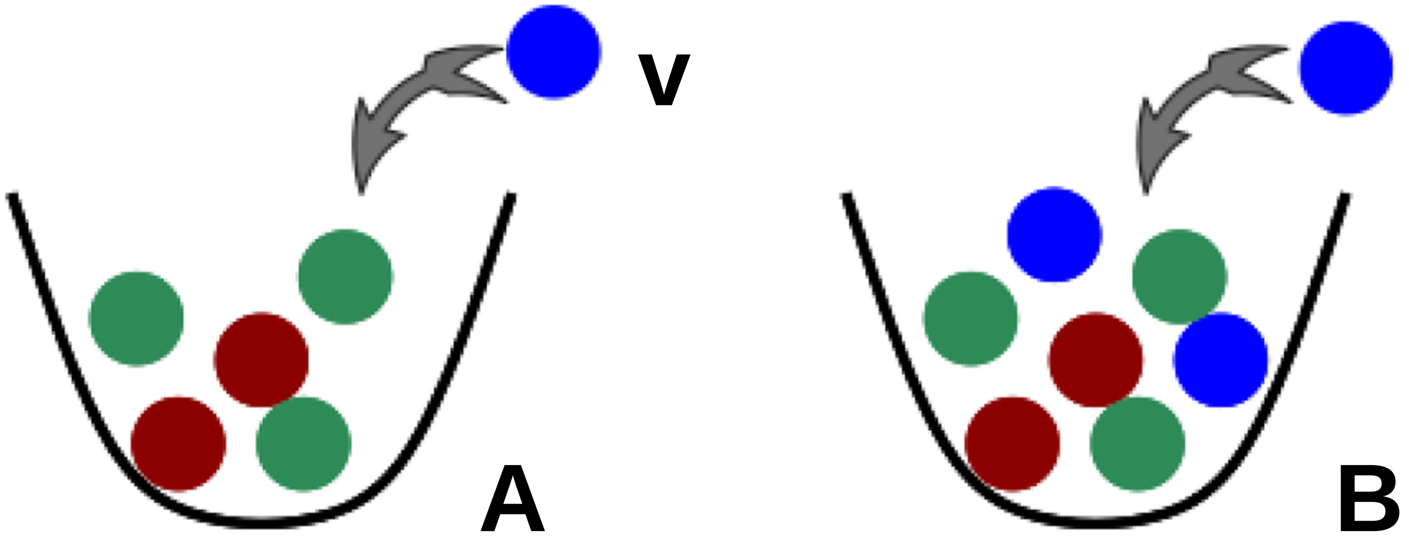


Guided Summarization



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 MI functions

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

(b) Instantiations of CG and CMI functions

CG	$f(\mathcal{A}\|\mathcal{P})$
FLCG	$\sum_{i\in\mathcal{V}} \max(\max_{j\in\mathcal{A}} S_{ij} - \max_{j\in\mathcal{P}} S_{ij}, 0)$
LOGDETCG	$\log\det(S_{\mathcal{A}} - \nu^2 S_{\mathcal{A},\mathcal{P}} S_{\mathcal{P}}^{-1} S_{\mathcal{A},\mathcal{P}}^T)$
GCCG	$f(\mathcal{A}) - 2\lambda\nu \sum_{i\in\mathcal{A}, j\in\mathcal{P}} S_{ij}$

CMI	$I_f(\mathcal{A};\mathcal{Q}\|\mathcal{P})$
FLCMI	$\sum_{i\in\mathcal{V}} \max(\min(\max_{j\in\mathcal{A}} S_{ij}, \max_{j\in\mathcal{Q}} S_{ij}) - \max_{j\in\mathcal{P}} S_{ij}, 0)$
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 (MI)

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

Submodular Conditional Gain (CG)

CG	$f(\mathcal{A}\|\mathcal{P})$
FLCG	$\sum_{i\in\mathcal{V}} \max(\max_{j\in\mathcal{A}} S_{ij} - \max_{j\in\mathcal{P}} S_{ij}, 0)$
LOGDETCG	$\log\det(S_{\mathcal{A}} - \nu^2 S_{\mathcal{A},\mathcal{P}} S_{\mathcal{P}}^{-1} S_{\mathcal{A},\mathcal{P}}^T)$
GCCG	$f(\mathcal{A}) - 2\lambda\nu \sum_{i\in\mathcal{A}, j\in\mathcal{P}} S_{ij}$

Submodular Conditional Mutual Information (CMI)

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

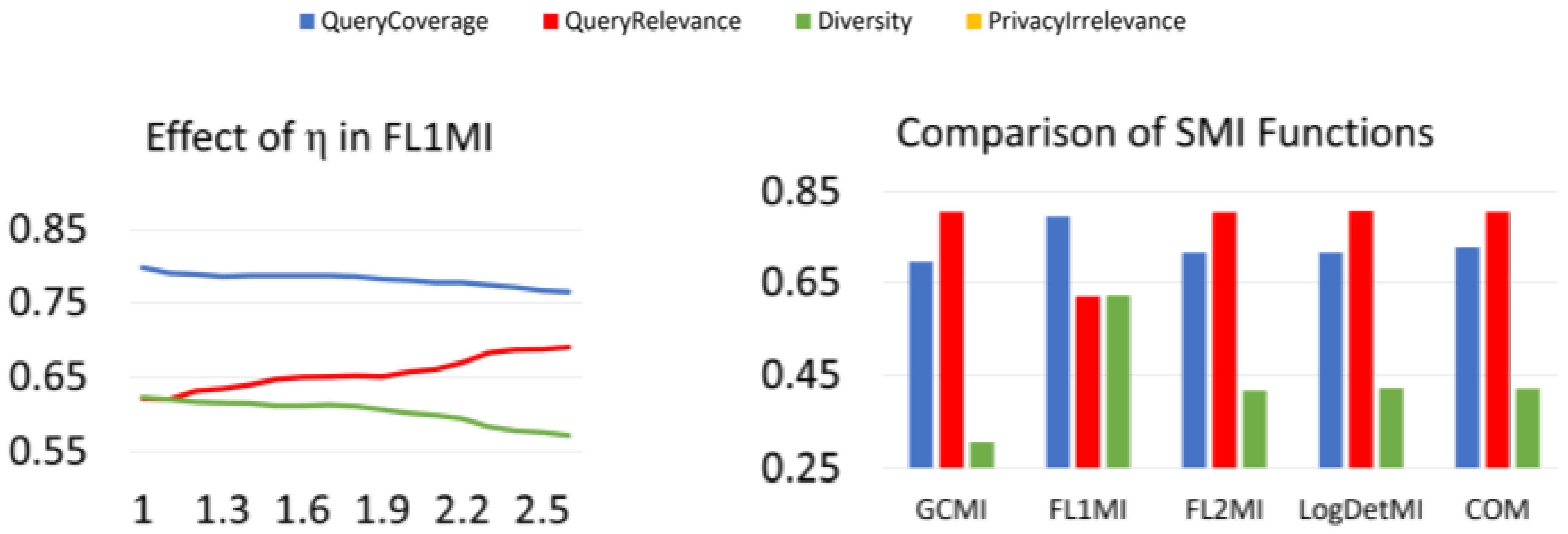
Guidance from an Auxiliary Set

- Guided subset selection requires the *guidance* to come from an auxiliary set V' that is held-out from the ground-set V.
- We define the submodular function on $\Omega = V \cup V'$
- The optimization problem is still defined on subsets $A \subseteq V$
- The query/private set can be a subset of V'.
- The optimization problem is then to maximize $I_f(A; Q)$ given a query set $Q \subseteq V'$, or $f(A|P)$ given a private set, $\mathrm{P} \subseteq V'$.

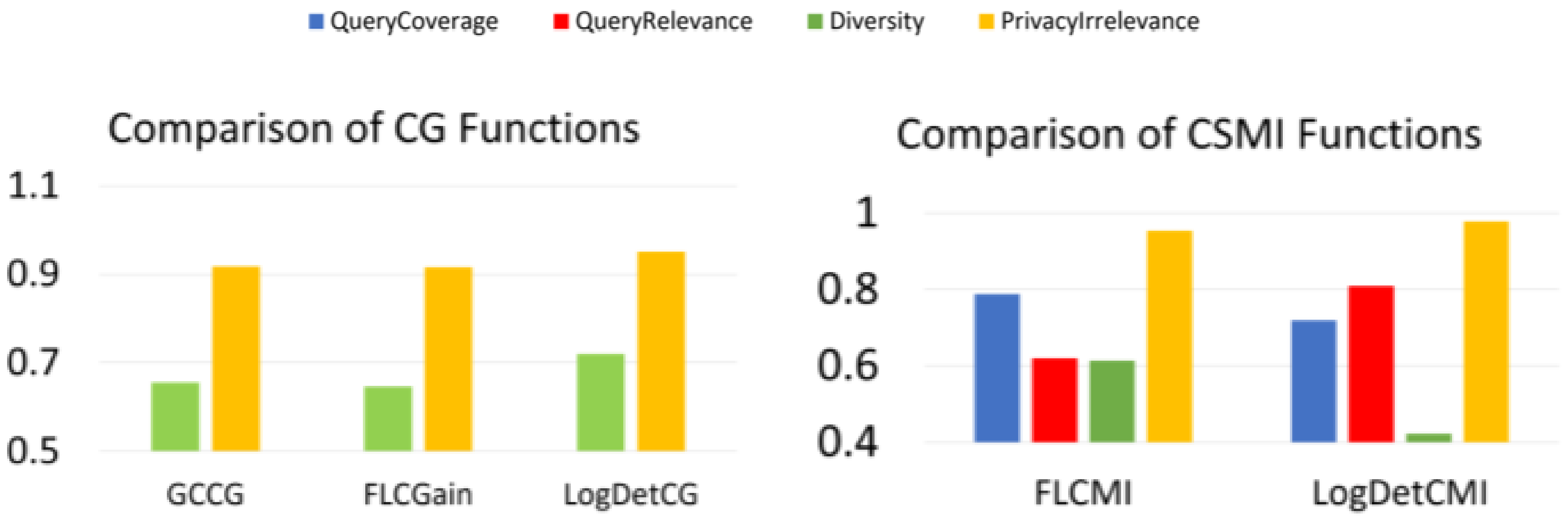
Modeling Semantics of PRISM

- We study characteristics of various PRISM instantiations with different parameters on synthetic datasets.
- We evaluate them based on the following characteristics:
- *Query-coverage* to be the fraction of queries covered by the subset.
- *Query-relevance* to be the fraction of the subset pertaining to the queries.
- *Diversity* to be the measure of how diverse are the points within the selected subset.
- *Privacy-irrelevance* to be the fraction of the subset not matching the private set.

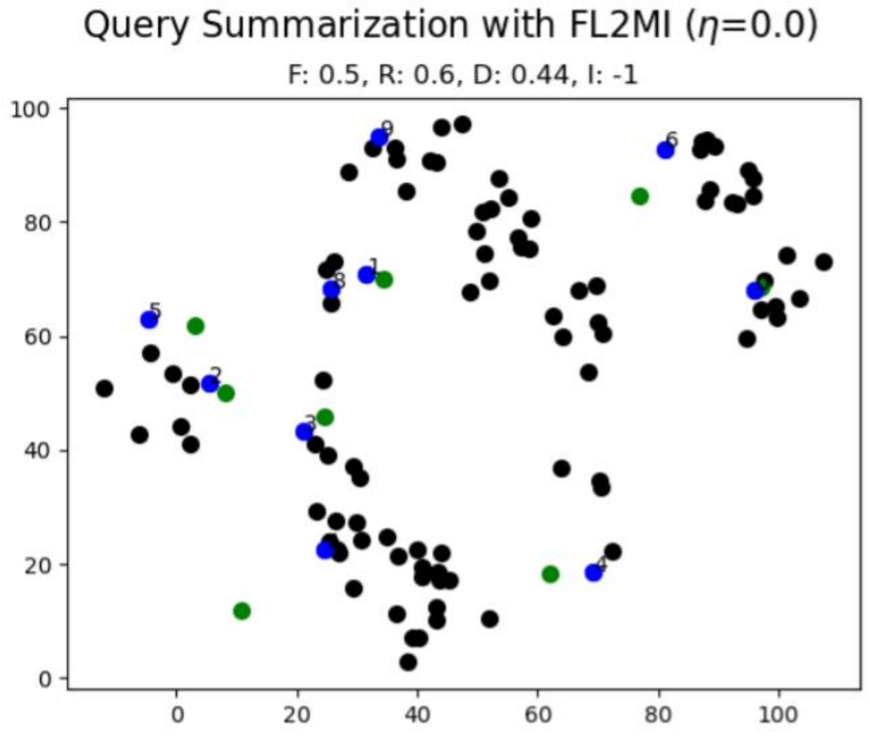
Modeling Semantics of PRISM



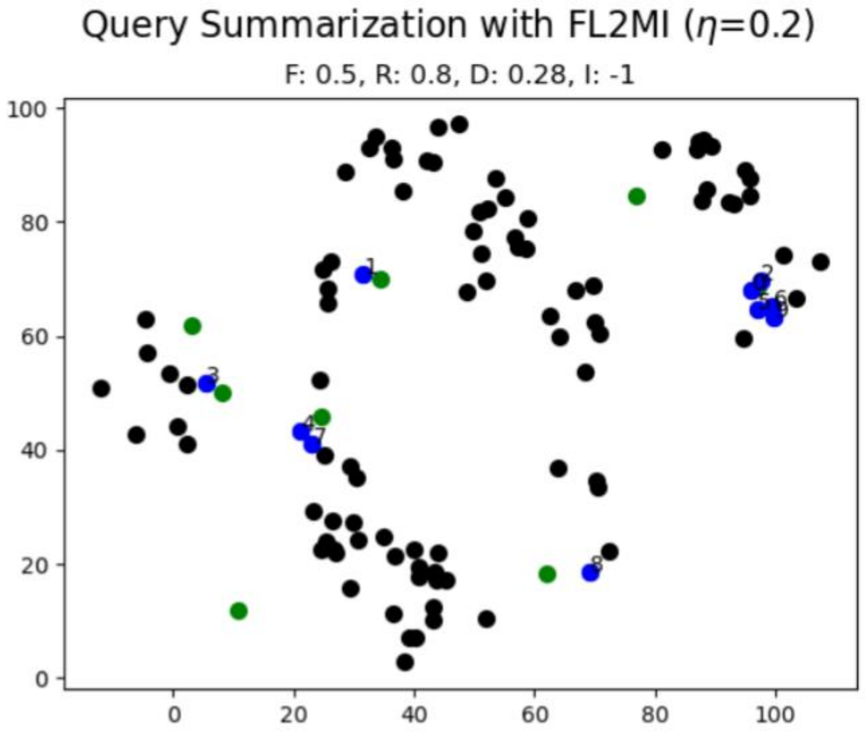
Modeling Semantics of PRISM



Modeling Semantics of PRISM



(a) $\eta = 0.0$



(b) $\eta = 0.2$

PRISM for Targeted Learning

Require: Initial Labeled set of Examples: $\mathcal{E}$, large unlabeled dataset: $\mathcal{U}$, A target subset/slice where we want to improve accuracy: $\mathcal{T}$, Loss function $\mathcal{L}$ for learning

1: Train model with loss $\mathcal{L}$ on labeled set $\mathcal{E}$ and obtain parameters θ_E

2: Compute the gradients $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{U}\}$ and $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{T}\}$.

3: Using the gradients, compute the similarity kernels and define a submodular function f and diversity function g

4: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A}\subseteq\mathcal{U}, |\mathcal{A}|\leq k} I_f(\mathcal{A}; T) + \gamma g(\mathcal{A})$

5: Obtain the labels of the elements in $\mathcal{A}^*$: $L(\hat{\mathcal{A}})$

6: Train a model on the combined labeled set $\mathcal{E} \cup L(\hat{\mathcal{A}})$

PRISM for Targeted Learning

Require: Initial Labeled set of Examples: $\mathcal{E}$, large unlabeled dataset: $\mathcal{U}$, A target subset/slice where we want to improve accuracy: $\mathcal{T}$, Loss function $\mathcal{L}$ for learning

1: Train model with loss $\mathcal{L}$ on labeled set $\mathcal{E}$ and obtain parameters θ_E

2: Compute the gradients $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{U}\}$ and $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{T}\}$.

3: Using the gradients, compute the similarity kernels and define a submodular function f and diversity function g

4: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq k} I_f(\mathcal{A}; T) + \gamma g(\mathcal{A})$

5: Obtain the labels of the elements in $\mathcal{A}^*$: $L(\hat{\mathcal{A}})$

6: Train a model on the combined labeled set $\mathcal{E} \cup L(\hat{\mathcal{A}})$

PRISM for Targeted Learning

Require: Initial Labeled set of Examples: $\mathcal{E}$, large unlabeled dataset: $\mathcal{U}$, A target subset/slice where we want to improve accuracy: $\mathcal{T}$, Loss function $\mathcal{L}$ for learning

1: Train model with loss $\mathcal{L}$ on labeled set $\mathcal{E}$ and obtain parameters θ_E

2: Compute the gradients $\{\nabla_{\theta_E} \mathcal{L}(x_i, y_i), i \in \mathcal{U}\}$ and $\{\nabla_{\theta_E} \mathcal{L}(x_i, y_i), i \in \mathcal{T}\}$.

3: Using the gradients, compute the similarity kernels and define a submodular function f and diversity function g

4: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq k} I_f(\mathcal{A}; T) + \gamma g(\mathcal{A})$

5: Obtain the labels of the elements in $\mathcal{A}^*$: $L(\hat{\mathcal{A}})$

6: Train a model on the combined labeled set $\mathcal{E} \cup L(\hat{\mathcal{A}})$

PRISM for Targeted Learning

Require: Initial Labeled set of Examples: $\mathcal{E}$, large unlabeled dataset: $\mathcal{U}$, A target subset/slice where we want to improve accuracy: $\mathcal{T}$, Loss function $\mathcal{L}$ for learning

1: Train model with loss $\mathcal{L}$ on labeled set $\mathcal{E}$ and obtain parameters θ_E

2: Compute the gradients $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{U}\}$ and $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{T}\}$.

3: Using the gradients, compute the similarity kernels and define a submodular function f and diversity function g

4: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A}\subseteq\mathcal{U}, |\mathcal{A}|\leq k} I_f(\mathcal{A}; T) + \gamma g(\mathcal{A})$

5: Obtain the labels of the elements in $\mathcal{A}^*$: $L(\hat{\mathcal{A}})$

6: Train a model on the combined labeled set $\mathcal{E} \cup L(\hat{\mathcal{A}})$

PRISM for Targeted Learning

Require: Initial Labeled set of Examples: $\mathcal{E}$, large unlabeled dataset: $\mathcal{U}$, A target subset/slice where we want to improve accuracy: $\mathcal{T}$, Loss function $\mathcal{L}$ for learning

1: Train model with loss $\mathcal{L}$ on labeled set $\mathcal{E}$ and obtain parameters θ_E

2: Compute the gradients $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{U}\}$ and $\{\nabla_{\theta_E}\mathcal{L}(x_i, y_i), i \in \mathcal{T}\}$.

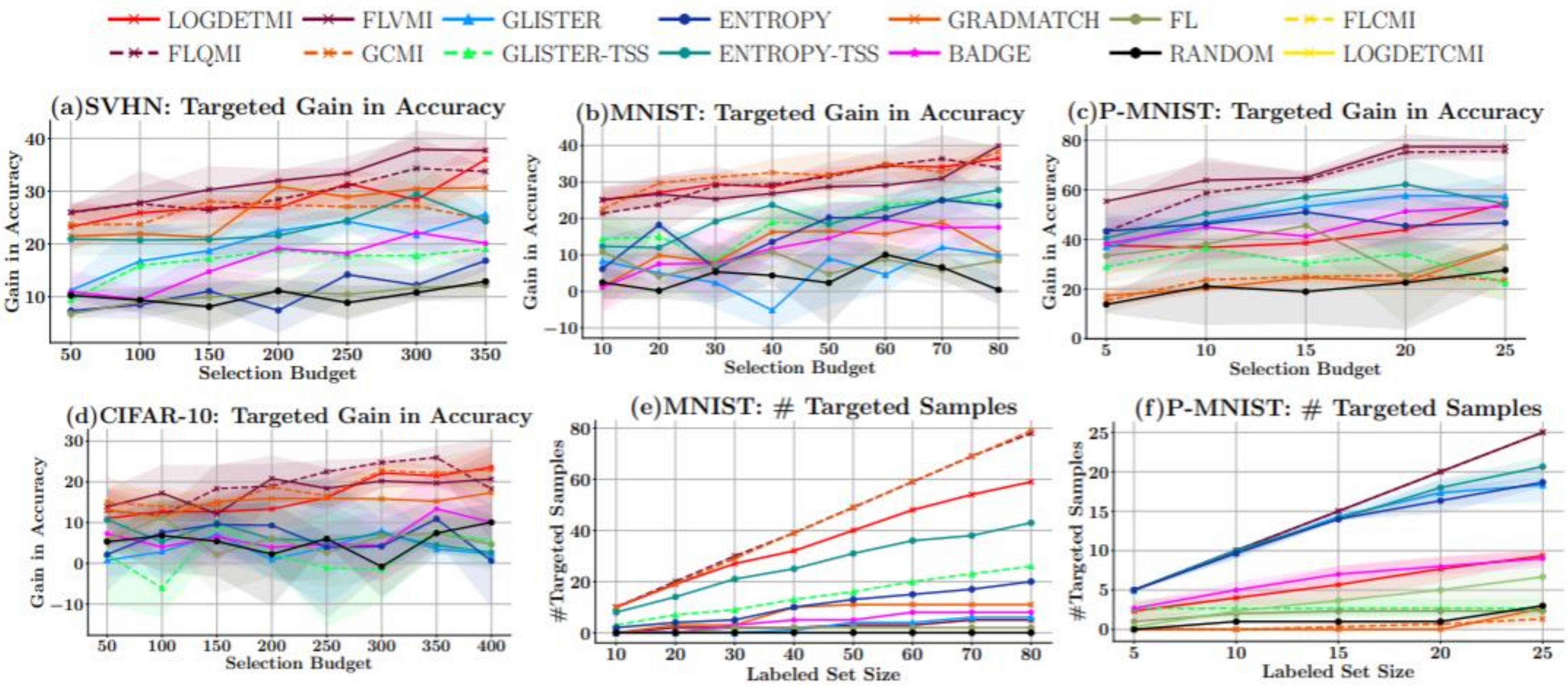
3: Using the gradients, compute the similarity kernels and define a submodular function f and diversity function g

4: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq k} I_f(\mathcal{A}; T) + \gamma g(\mathcal{A})$

5: Obtain the labels of the elements in $\mathcal{A}^*$: $L(\hat{\mathcal{A}})$

6: Train a model on the combined labeled set $\mathcal{E} \cup L(\hat{\mathcal{A}})$

Results – Targeted Learning



MI based functions consistently outperform all baselines by ~ 20 − 30% in terms of average accuracy on target classes.

PRISM's Unified Framework for Guided Summarization

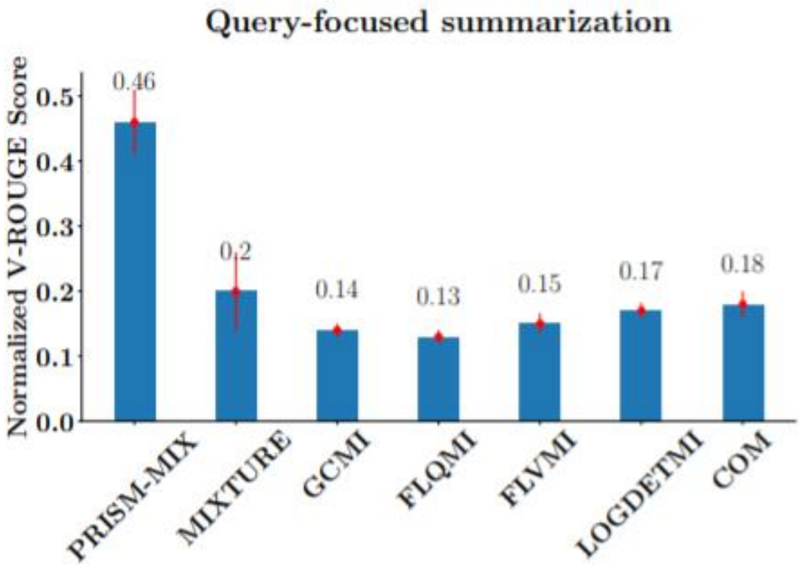
$$\max_{A:A\subseteq k} I_f(A;Q|P)$$

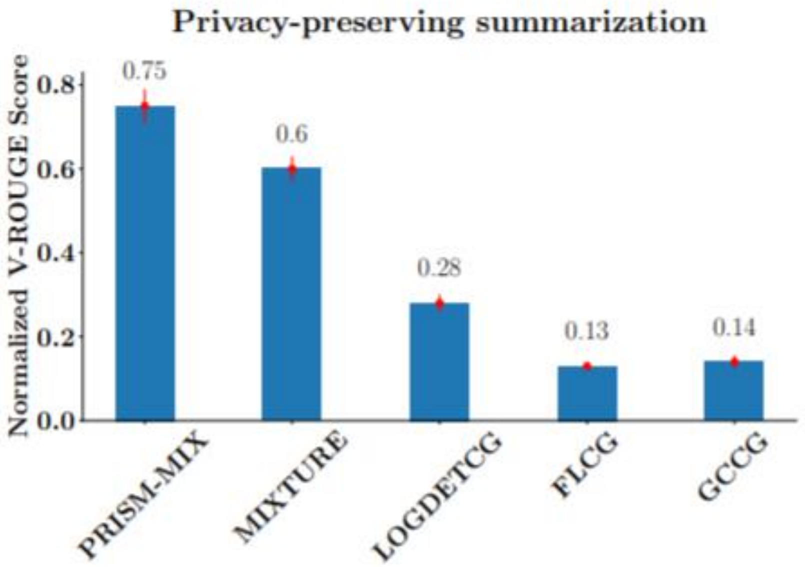
- Optimizing the CMI function can be viewed as a master optimization problem for multiple summarization tasks.
- *Generic summarization:* $Q \leftarrow V, P \leftarrow \emptyset$
- *Query-focused summarization:* $Q \leftarrow Q, P \leftarrow \emptyset$
- *Privacy-preserving summarization :* $Q \leftarrow \emptyset, P \leftarrow P$
- *Query-focused and Privacy-preserving summarization :* $Q \leftarrow Q, P \leftarrow P$

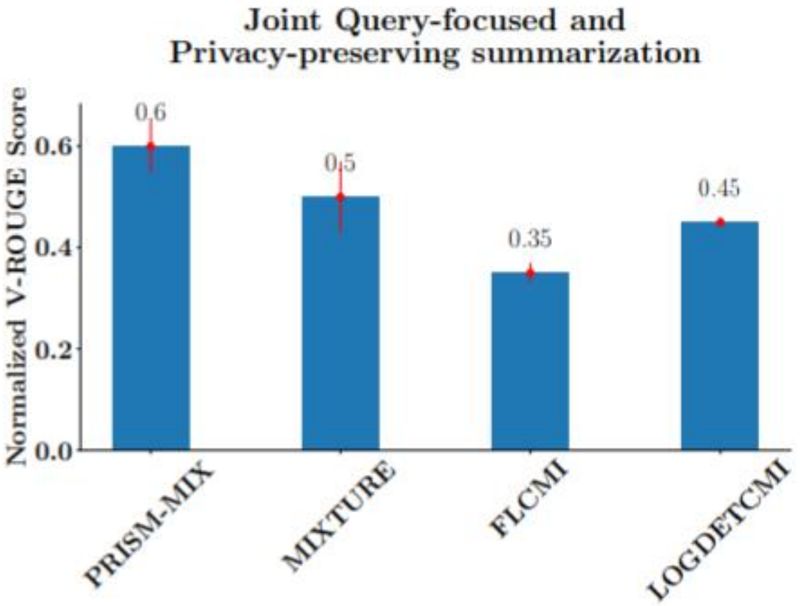
Parameter Learning in PRISM for Guided Summarization

- For guided summarization, we learn a mixture of PRISM functions (PRISM-MIX) where the weights and internal parameters are jointly learned.
- The mixture is learned using a max-margin formulation supervised by summaries generated by humans.
- For generic summarization, we add the standard submodular functions modeling representation, diversity, coverage.
- For query-focused summarization and privacy-preserving summarization, we instead use the MI and CG versions of the PRISM functions.
- During inference, we instantiate the mixture model with the learned parameters and maximize it to get the desired summaries.

Results – Guided Summarization







- Dataset with 14 image collections with 100 images each, and 50-250 human summaries per collection.
- We compare PRISM-MIX with individual components used in the mixture.
- MIXTURE model uses the same components as PRISM-MIX without learning the internal parameters of PRISM functions.

Conclusion

- We presented PRISM, a rich class of functions for guided subset selection.
- PRISM allows to model a broad spectrum of semantics across query-relevance, diversity, query-coverage and privacy-irrelevance.
- We demonstrated its effectiveness in targeted learning as well as in guided summarization.
- In our paper, we showed that PRISM has interesting connections to several past work, further reinforcing its utility.
- Through experiments on targeted learning and guided summarization for diverse datasets, we empirically verified the superiority of PRISM over existing methods.

Thank You

For more details, do visit our poster.