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

Suraj Kothawade[1], Vishal Kaushal [2], Ganesh Ramakrishnan [2], Jeff Bilmes [3], Rishabh Iyer [1,2]

[1] University of Texas at Dallas
[2] Indian Institute of Technology, Bombay
[3] University of Washington, Seattle
suraj.kothawade@utdallas.edu, vkaushal@cse.iitb.ac.in, ganesh@cse.iitb.ac.in, bilmes@uw.edu, rishabh.iyer@utdallas.edu

Abstract

With ever-increasing dataset sizes, subset selection techniques are becoming increasingly important for a plethora of tasks. It is often necessary to *guide* the subset selection to achieve certain desiderata, which includes *focusing or targeting* certain data points, while *avoiding* others. Examples of such problems include: i) *targeted learning*, where the goal is to find subsets with rare classes or rare attributes on which the model is underperforming, and ii) *guided summarization*, where data (*e.g.*, image collection, text, document or video) is summarized for quicker human consumption with specific additional user intent. Motivated by such applications, we present PRISM, a rich class of PaRameterIzed Submodular information Measures. Through novel functions and their parameterizations, PRISM offers a variety of modeling capabilities that enable a trade-off between desired qualities of a subset like diversity or representation and similarity/dissimilarity with a set of data points. We demonstrate how PRISM can be applied to the two real-world problems mentioned above, which require guided subset selection. In doing so, we show that PRISM interestingly generalizes some past work, therein reinforcing its broad utility. Through extensive experiments on diverse datasets, we demonstrate the superiority of PRISM over the state-of-the-art in targeted learning and in guided image-collection summarization.

1 Introduction

Recent times have seen explosive growth in data across several modalities, including text, images, and videos. This has given rise to the need for finding techniques for selecting effective smaller data subsets with specific characteristics for a variety of down-stream tasks. Often, we would like to *guide* the data selection to either *target* or *avoid* a certain set of data slices. One application is, what we call, *targeted learning*, where the goal is to select data points similar to data slices on which the model is currently performing poorly. These slices are data points that either belong to rare classes or have common rare attributes (*e.g.*, color, background, *etc.*). An example of such a scenario is shown in Fig. 1(a), where a self-driving car model struggles in detecting "cars in a dark background" because of a lack of such images in the training set. The targeted learning problem is to augment the training dataset with more of such rare images, with an aim to improve model performance. Another example is detecting cancers in biomedical imaging datasets, where the number of cancerous images are often a small fraction of the non-cancerous images.

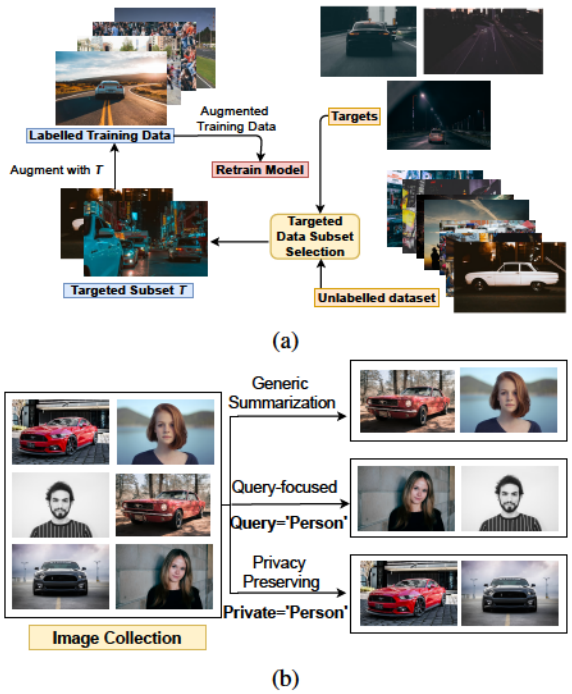

Figure 1: Applications of guided subset selection. (a) *Targeted learning:* improving a model's performance on night images (target), which are under-represented in the training data. This is achieved by augmenting it with a subset matching the target. (b) *Guided summarization:* finding a summary similar to a query set or a summary dissimilar to a private set.

Another application comes from the summarization task, where an image collection, a video, or a text document is summarized for quicker human consumption by eliminating redundancy, while preserving the main content. While a number of applications require *generic* summarization (*i.e.*, simply picking a representative and diverse subset of the massive dataset), it is often important to capture certain user intent in summarization. We call this *guided summarization*. Examples of guided summarization include: (i) *query-focused summarization* (Sharghi, Gong, and Shah 2016; Xiao et al. 2020), where a summary similar to a specific query is desired,

and (ii) *privacy-preserving summarization*, where a summary dissimilar to a given private set of data points is desired (say, for privacy issues). See Fig. 1(b) for a pictorial illustration.

1.1 Our Contributions

PRISM Framework: We define PRISM through different instantiations and parameterizations of various submodular information measures (Sec. 2). These allow for modeling a spectrum of semantics required for guided subset selection, like relevance to a query set, irrelevance to a private set, and diversity among selected data points. We study the effect of parameter trade-off among these different semantics and present interesting insights.

PRISM for Targeted Learning: We present a novel algorithm (Sec. 3.1, Algo. 1) to apply PRISM for targeted learning, which aims to improve a model's performance on rare slices of data. Specifically, we show that submodular information measures are very effective in finding the examples from the rare classes in a large unlabeled set (akin to finding a needle in a haystack). On several image classification tasks, PRISM obtains $\approx$ 20-30% gain in accuracy of rare classes ($\approx$ 12% more than existing approaches) by just adding a few additional labeled points from the rare classes. Furthermore, we show that PRISM is 20$\times$ to 50$\times$ more label-efficient compared to random sampling, and 2$\times$ to 4$\times$ more label-efficient compared to existing approaches (see Sec. 4.1). We also show that Algo. 1 generalizes some existing approaches for data subset selection, reinforcing its utility (Sec. 3.3).

PRISM for Guided Summarization. We propose a learning framework for guided summarization using PRISM (Sec. 3.2). We show that PRISM offers a *unified* treatment to the different flavors of guided summarization (query-focused and privacy-preserving) and generalizes some existing approaches to summarization, again reinforcing its utility. We show that it outperforms other existing approaches on a real-world image collections dataset (Sec. 4.2).

1.2 Related Work

Submodularity and Submodular Information Measures: Submodularity (Fujishige 2005) is a rich yet tractable subfield of non-linear combinatorial optimization (Krause and Golovin 2014).We provide novel formulations of the recently introduced class of submodular information measures (Gupta and Levin 2020; Iyer et al. 2021) for guided data subset selection.

Data Subset Selection, Coresets, and Active Learning: A number of papers have studied data subset selection in different applications and settings. Several recent papers have studied data subset selection for speeding up training. These include approaches involving submodularity (Wei, Iyer, and Bilmes 2015; Kaushal et al. 2019a), gradient coresets (Mirzasoleiman, Bilmes, and Leskovec 2020; Killamsetty et al. 2021) and bi-level based coresets (Killamsetty et al. 2020). Another application is active learning, where the goal is to select and label a subset of unlabeled data points to improve model performance (Settles 2009). Several recent approaches which combine notions of diversity and uncertainty have become popular (Wei, Iyer, and Bilmes 2015; Sener and Savarese 2018; Ash et al. 2020). One such state-of-the-art approach is BADGE (Ash et al. 2020), which samples points that have diverse hypothesized gradients. Most of these paradigms have been studied in the setting of generic data subset selection, and are ineffective when it comes to guided subsets. Some recent works like GRAD-MATCH (Killamsetty et al. 2021) and GLISTER (Killamsetty et al. 2020) select subsets based on a held out validation set, which can be a rare slice of data. Similarly, (Kirchhoff and Bilmes 2014) compute a targeted subset of training data in the spirit of transductive learning for machine translation.

Summarization: A number of instances of summarization have been studied in the past, including image collection summarization (Celis and Keswani 2020; Ozkose et al. 2019; Singh, Virmani, and Subramanyam 2019; Tschiatschek et al. 2014), text/document summarization (Lin and Bilmes 2012; Chali, Tanvee, and Nayeem 2017; Yao, Wan, and Xiao 2017), and video summarization (Kaushal et al. 2019c,b; Gygli, Grabner, and Gool 2015; Ji et al. 2019). While most of these works have focused on generic summarization, some have also studied query-focused video summarization (Sharghi, Gong, and Shah 2016; Sharghi, Laurel, and Gong 2017; Vasudevan et al. 2017; Xiao et al. 2020; Jiang and Han 2019), and query-focused document summarization (Lin and Bilmes 2011; Li, Li, and Li 2012). To the best of our knowledge, PRISM is the first attempt to offer a unified treatment to the different flavors of summarization.

2 The PRISM Framework

2.1 Preliminaries

Submodular functions: Let $\mathcal{V} = \{1, 2, 3, ..., n\}$ denote the *ground-set* and f denote a set function $f : 2^{\mathcal{V}} \rightarrow \Re$. The function f is *submodular* if it satisfies the diminishing marginal returns property; namely $f(j|\mathcal{X}) \geq f(j|\mathcal{Y}), \forall \mathcal{X} \subseteq \mathcal{Y} \subseteq \mathcal{V}, j \notin \mathcal{Y}$ (Fujishige 2005). Submodularity (along with monotonicity) ensures that a greedy algorithm achieves a $1 - 1/e$ approximation when f is maximized (Nemhauser, Wolsey, and Fisher 1978).

Submodular Conditional Gain (CG): Given sets $\mathcal{A}, \mathcal{P} \subseteq \mathcal{V}$, the CG, $f(\mathcal{A}|\mathcal{P})$, is the gain in function value by adding $\mathcal{A}$ to $\mathcal{P}$. Thus $f(\mathcal{A}|\mathcal{P}) = f(\mathcal{A} \cup \mathcal{P}) - f(\mathcal{P})$. Intuitively, $f(\mathcal{A}|\mathcal{P})$ measures how different $\mathcal{A}$ is from $\mathcal{P}$, where $\mathcal{P}$ is the *conditioning set* or the *private set*.

Submodular Mutual Information (MI): Given sets $\mathcal{A}, \mathcal{Q} \subseteq \mathcal{V}$, the MI (Gupta and Levin 2020; Iyer et al. 2021) is defined as $I_f(\mathcal{A}; \mathcal{Q}) = f(\mathcal{A}) + f(\mathcal{Q}) - f(\mathcal{A} \cup \mathcal{Q})$. Intuitively, this measures the similarity between $\mathcal{Q}$ and $\mathcal{A}$ where $\mathcal{Q}$ is the query set.

Submodular Conditional Mutual Information (CMI): CMI is defined using CG and MI as $I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P}) = f(\mathcal{A} \cup \mathcal{P}) + f(\mathcal{Q} \cup \mathcal{P}) - f(\mathcal{A} \cup \mathcal{Q} \cup \mathcal{P}) - f(\mathcal{P})$. Intuitively, CMI jointly models the mutual similarity between $\mathcal{A}$ and $\mathcal{Q}$ and their collective dissimilarity from $\mathcal{P}$.

Properties: CG, MI, and CMI are non-negative and monotone in one argument with the other fixed (Gupta and Levin 2020; Iyer et al. 2021). CMI and MI are not necessarily submodular in one argument (with the others fixed) (Iyer et al. 2021). However, several of the instantiations we define below turn out to be submodular. With this background, we present our unique and novel formulations, leading to PRISM.

2.2 Guidance from an Auxiliary Set

We formulate the above submodular information measures to handle the case when the *guidance* can come from an auxiliary set $\mathcal{V}'$ different from the ground set $\mathcal{V}$ – a requirement common in several guided subset selection tasks. Let $\Omega = \mathcal{V} \cup \mathcal{V}'$. We define a set function $f : 2^{\Omega} \rightarrow \Re$. Although f is defined on Ω, the discrete optimization problem will only be defined on subsets $\mathcal{A} \subseteq \mathcal{V}$. To find an optimal subset (i) given a query set $\mathcal{Q} \subseteq \mathcal{V}'$, we can define $g_{\mathcal{Q}}(\mathcal{A}) = I_f(\mathcal{A}; \mathcal{Q})$, $\mathcal{A} \subseteq \mathcal{V}$ and maximize the same; (ii) given a private set $\mathcal{P} \subseteq \mathcal{V}'$, we can define $h_{\mathcal{P}}(\mathcal{A}) = f(\mathcal{A}|\mathcal{P})$, $\mathcal{A} \subseteq \mathcal{V}$, as the function to be maximized.

2.3 Restricted Submodularity to Enable a Richer Class of MI and CG Functions

While submodular functions are expressive, many natural choices are not submodular everywhere. We do not need f to be submodular everywhere on Ω, since the sets we are optimizing on, are subsets of $\mathcal{V}$. Instead of requiring the submodular inequality to hold for all pairs of sets $(\mathcal{X}, \mathcal{Y}) \in 2^{\Omega} \times 2^{\Omega}$, we can consider only subsets of this power set pivoting on $\mathcal{V} \subseteq \Omega$. In particular, define a subset $\mathcal{C} \subseteq 2^{\Omega}$. Then *restricted submodularity* on $\mathcal{C}$ satisfies $f(\mathcal{X}) + f(\mathcal{Y}) \geq f(\mathcal{X} \cup \mathcal{Y}) + f(\mathcal{X} \cap \mathcal{Y}), \forall (\mathcal{X}, \mathcal{Y}) \in \mathcal{C}$. Instances of restricted submodularity in the form of intersecting and crossing submodular functions have been considered in the past (Fujishige 2005). We consider the following form of restricted submodularity. Given sets $\mathcal{V}$ and $\mathcal{V}'$ as above, define $\mathcal{C}(\mathcal{V}, \mathcal{V}') \subseteq 2^{\Omega}$ to be such that the sets $(\mathcal{X}, \mathcal{Y}) \in \mathcal{C}(\mathcal{V}, \mathcal{V}')$ satisfy either of the following conditions: i) $\mathcal{X} \subseteq \mathcal{V}$ or $\mathcal{X} \subseteq \mathcal{V}'$ and $\mathcal{Y}$ is *any* set, or ii) $\mathcal{X}$ is *any* set and $\mathcal{Y} \subseteq \mathcal{V}$ or $\mathcal{Y} \subseteq \mathcal{V}'$. We call the MI of a restricted submodular function as Generalized Mutual Information function (GMI). We use this notion of GMI to define Concave Over Modular (COM). The GMI function is non-negative and monotone (see Appendix B.1).

2.4 Instantiations & Parameterizations in PRISM

In this section, we discuss the expressions for different instantiations of the above measures using different functions. We refer to them as •MI or •CG or •CMI where • is the submodular function using which the respective MI, CG or CMI measure is instantiated. While different submodular functions naturally model different characteristics such as representation, coverage, *etc.* (Kaushal et al. 2019c,b), the instantiations presented here additionally model similarity and dissimilarity to query and private sets respectively. These instantiations have parameters λ, η and/or ν, that govern the interplay among different characteristics. In several instantiations, we invoke a similarity matrix S where S_{ij} measures the similarity between elements i and j of sets that will be correspondingly specified. *The rich class of functions in* PRISM *thus helps model a broad spectrum of semantics.* The mathematical expressions for each function are summarized in Tab. 1. Below, we provide further notations and intuitions for using these functions.

Log Determinant (LogDet): Let $S_{\mathcal{A},\mathcal{Q}}$ be the cross-similarity matrix between the items in sets $\mathcal{A}$ and $\mathcal{Q}$. We construct a similarity matrix $S^{\eta,\nu}$ (on a base matrix S) in such a way that the cross-similarity between $\mathcal{A}$ and $\mathcal{Q}$ is multiplied by η (*i.e.*, $S^{\eta,\nu}_{\mathcal{A},\mathcal{Q}} = \eta S_{\mathcal{A},\mathcal{Q}}$) to control the trade-off between query-relevance and diversity. Similarly, the cross-similarity between $\mathcal{A}$ and $\mathcal{P}$ by ν (*i.e.*, $S^{\eta,\nu}_{\mathcal{A},\mathcal{P}} = \nu S_{\mathcal{A},\mathcal{P}}$) to control the strictness of privacy constraints. Higher values of ν ensure stricter privacy constraints, such as in the context of privacy-preserving summarization, by tightening the extent of dissimilarity of the subset from the private set. Given the standard form of LogDet as $f(\mathcal{A}) = \log\det(S^{\eta,\nu}_{\mathcal{A}})$, we provide the PRISM expressions in Tab. 1. For simplicity of notation, CMI is presented with $\nu = \eta = 1$. We defer the derivation and the general expression for CMI to the Appendix C.1.

Facility Location (FL): We introduce two variants of the MI functions for the FL function which is defined as: $f(\mathcal{A}) = \sum_{i \in \Omega} \max_{j \in \mathcal{A}} S_{ij}$. The first variant is defined over $\mathcal{V}$ (FLVMI) (Iyer et al. 2021), in Tab. 1(a). We derive another variant defined over $\mathcal{Q}$ (FL**Q**MI) which considers only cross-similarities between data points and the target. This MI expression has interesting characteristics different from those of FLVMI. In particular, whereas FLVMI gets saturated (*i.e.*, once the query is satisfied, there is no gain in picking another query-relevant data point), FLQMI just models the pairwise similarities of target to data points and vice versa. Moreover, FLQMI only requires a $\mathcal{Q} \times \mathcal{V}$ kernel, which makes it very efficient to optimize. We provide the expressions in Tab. 1 and defer the derivations to Appendix C.2. We multiply the similarity kernel S used in MI and CG expressions of FL by η and ν as done in the case of LogDet.

Concave Over Modular (COM): The notion of GMI functions (Sec. 2.3) allows us to characterize a rich class of concave over modular functions as GMI functions. Define a set function $f_\eta(\mathcal{A})$ as: $f_\eta(\mathcal{A}) = \eta \sum_{i \in \mathcal{V}'} \max(\psi(\sum_{j \in \mathcal{A} \cap \mathcal{V}} S_{ij}), \psi(\sqrt{n} \sum_{j \in \mathcal{A} \cap \mathcal{V}'} S_{ij}))$ $+ \sum_{i \in \mathcal{V}} \max(\psi(\sum_{j \in \mathcal{A} \cap \mathcal{V}'} S_{ij}), \psi(\sqrt{n} \sum_{j \in \mathcal{A} \cap \mathcal{V}} S_{ij}))$, where ψ is a concave function and $f_\eta(\mathcal{A})$ is restricted submodular. We state the expression for its GMI function in Tab. 1(a) and provide the derivation in Appendix B.2.

Graph Cut (GC): The GC function is defined as $f(\mathcal{A}) = \sum_{i \in \mathcal{A}, j \in \mathcal{V}} S_{ij} - \lambda \sum_{i,j \in \mathcal{A}} S_{ij}$. The parameter λ captures the trade-off between diversity and representativeness. The PRISM expressions of GC are presented in Tab. 1. Note that the CMI expression for GC is not useful as it does not involve the private set and is exactly the same as the MI version (proof in Appendix C.3). Like in the LogDet case, we introduce an additional parameter ν in GCCG to control the strictness of privacy constraints. Again, this is easily modeled in the GC objective by multiplying the cross-similarity between data points and the private instances by ν.

Computational Complexity: In terms of compute complexity, GCMI and FLQMI are linear in $|\mathcal{V}|$ (since $\mathcal{Q}$ is typically small). However, FLVMI and LOGDETMI are quadratic in the size of the unlabeled set due to requiring the kernel. Hence, for massive datasets, GCMI and FLQMI are preferable (more details in Appendix D). We also discuss (in Appendix D) a partitioning based approach where we divide the datasets into smaller partitions and run the $\mathcal{V} \times \mathcal{V}$ kernel based functions (FLVMI and LOGDETMI) on the individual partitions, thereby making them more scalable.

Table 1: Instantiations of PRISM. Note that the functions formulate similarity with query set Q and dissimilarity with private set P which are the building blocks for targeted data subset selection.

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

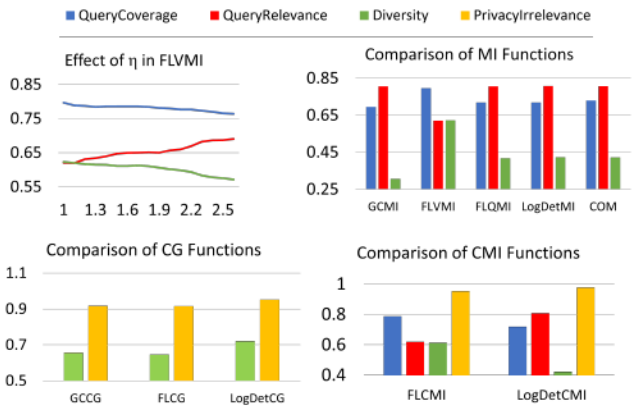

Figure 2: Behavior of different functions in PRISM and effect of parameters. All plots share the legend.

2.5 Modeling Semantics of PRISM

To empirically verify the intuitive understanding of the expressions, we maximize the different functions in PRISM individually on a synthetically created dataset with different parameters, and study the characteristics of the subsets qualitatively and quantitatively. For evaluation, we define *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, and *privacy-irrelevance* to be the fraction of the subset *not* matching the private set. We present representative results in Fig. 2 and defer detailed results to Appendix E. For MI functions, we verify that increasing η tends to increase query-relevance while reducing query-coverage and diversity (top-left, Fig. 2). Also, while GCMI lies at one end of the spectrum favoring query-relevance, FLVMI lies at the other end favoring diversity and query coverage. FLQMI, LOGDETMI and COM lie somewhere in between (top-right, Fig. 2). As expected, increasing ν increases privacy-irrelevance for CG functions. We also observe that LOGDETCG outperforms FLCG and GCCG both in terms of diversity and privacy-irrelevance (bottom-left, Fig. 2). For CMI functions, we see that FLCMI tends to favor query-coverage and diversity in contrast to query-relevance and privacy-irrelevance, while LOGDETCMI favors query-relevance and privacy-irrelevance over query-coverage and diversity (bottom-right, Fig. 2).

3 PRISM for Guided Subset Selection

In this section, we discuss the use of PRISM for guided data subset selection and illustrate its utility for targeted learning and guided summarization.

3.1 Targeted Learning

We first apply PRISM to *targeted learning*, where the goal is to improve a model's accuracy on some target classes at a given additional labeling cost and without compromising on the overall accuracy (see Fig. 1(a)). Let $\mathcal{E}$ be an initial training set of labeled instances, $\mathcal{T}$ be the target set of examples on which the user desires good performance, $\mathcal{P}$ be the private set of examples that the user wants to avoid, and $\mathcal{U}$ be a large unlabeled dataset. We maximize a CMI function $I_f(\mathcal{A};\mathcal{T}|\mathcal{P})$ towards computing an optimal subset $\hat{\mathcal{A}} \subseteq \mathcal{U}$ of size k *similar* to $\mathcal{T}$ and *dissimilar* to $\mathcal{P}$. Note that when $\mathcal{P} = \emptyset$, CMI is equivalent to MI. We then augment $\mathcal{E}$ with labeled $\hat{\mathcal{A}}$ and re-train the model. Through the aforementioned (Sec. 2), the diverse class of MI functions in PRISM offers a natural and effective approach for targeted subset selection by using the query set $\mathcal{Q}$ as the target set $\mathcal{T}$. The approach is outlined in Algo. 1. Similar to (Ash et al. 2020; Killamsetty et al. 2020, 2021; Mirzasoleiman, Bilmes, and Leskovec 2020), we use last-layer gradients of the model to represent the data points and use them to compute the similarity kernel S. Specifically, we define pairwise similarities $S_{ij} = \langle \nabla_\theta \mathcal{L}_i(\theta), \nabla_\theta \mathcal{L}_j(\theta) \rangle$, where $\mathcal{L}_i(\theta) = \mathcal{L}(x_i, y_i, \theta)$ is the loss on the ith data point and θ denotes model parameters. Note that the target need not correspond to class(es) but could be any attribute of data that the user is interested in. For instance, the target could be mining images with people at night (here *night* is an attribute).

3.2 Guided Summarization

Next, we apply PRISM for *guided summarization*. In this task, we are given a set $\mathcal{V}$ of data points (images, sentences of a document, or frames/shots of a video), and the goal is to find a summary $\mathcal{A} \subseteq \mathcal{V}$ with certain desired characteristics. In query-focused summarization, we find a summary that is semantically similar to the query set, while in privacy-preserving summarization, the obtained summary should *not* contain data points that are similar to the private set.

Algorithm 1: Application of PRISM for **Targeted Learning**

Require: $\mathcal{E}$: initial labeled set, $\mathcal{U}$: large unlabeled set, $\mathcal{T}$: a target subset/slice, $\mathcal{P}$: a set to be avoided, k: selection budget, $\mathcal{L}$: loss function

1: Train model with loss $\mathcal{L}$ on labeled set $\mathcal{E}$ to obtain model parameters $\theta_{\mathcal{E}}$ {Obtain initial accuracy}
2: Compute the gradients $\{\nabla_{\theta_{\mathcal{E}}}\mathcal{L}(x_i, y_i), i \in \mathcal{U}\}$ (using hypothesized labels) and $\{\nabla_{\theta_{\mathcal{E}}}\mathcal{L}(x_i, y_i), i \in \mathcal{T}\}$ {Obtain vectors for computing kernel in Step 3}
3: Compute the similarity kernels S and define a CMI function $I_f(\mathcal{A}; \mathcal{T}|\mathcal{P})$ {Instantiate Functions}
4: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A}\subseteq\mathcal{U},|\mathcal{A}|\leq k}(I_f(\mathcal{A}; \mathcal{T}|\mathcal{P})$ {Obtain the subset optimally matching the target}
5: Obtain the labels of the elements in $\hat{\mathcal{A}}$ as $L(\hat{\mathcal{A}})$ {Procure labels on the selected subset}
6: Train a model on the combined labeled set $\mathcal{E} \cup L(\hat{\mathcal{A}})$ {Augment training data}

PRISM's Unified Framework for Guided Summarization: Given sets $\mathcal{Q}$ and $\mathcal{P}$, and a submodular function f, consider the following *master optimization problem* involving CMI: $\max_{\mathcal{A}:|\mathcal{A}|\leq k} I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P})$. We discuss how the different flavors of summarization can be seen as special cases of this master optimization problem. Setting $\mathcal{Q} \leftarrow \mathcal{V}$ and $\mathcal{P} \leftarrow \emptyset$ yields generic summarization. Similarly, setting $\mathcal{Q} \leftarrow \mathcal{Q}$ and $\mathcal{P} \leftarrow \emptyset$ yields query-focused summarization with a query-set $\mathcal{Q}$. Setting $\mathcal{Q} \leftarrow \mathcal{V}$ and $\mathcal{P} \leftarrow \mathcal{P}$ gives privacy-preserving summarization. This framework allows us to address yet another flavor: *joint query-focused and privacy preserving summarization* where we set $\mathcal{Q} \leftarrow \mathcal{Q}$ and $\mathcal{P} \leftarrow \mathcal{P}$.

Parameter Learning in PRISM for Guided Summarization: As discussed in Sec. 2, different instantiations of PRISM along with their parameters offer a wide spectrum of modeling characteristics. Hence, when used individually, each imparts certain characteristics to the summaries. For summarization, we thus propose learning a mixture model supervised by summaries generated by humans. We learn a mixture of PRISM functions (Lin and Bilmes 2012; Kaushal et al. 2019c,b; Tschiatschek et al. 2014) where the weights and the internal parameters λ, ν, η of the functions are jointly learned. We denote our parameter vector by $\Theta = (w, \eta, \lambda, \nu)$, and our PRISM mixture model by $F(\Theta) = \sum_i w_i f_i(\mathcal{A}, \gamma, \eta, \nu)$, with each f_i being either one of the functions in PRISM or one of pure diversity and representation functions such as Disparity-Sum and FL. Then, given N training examples, $(\mathcal{V}^{(n)}, \mathcal{Y}^{(n)})_{n=1}^N$ we apply gradient descent to learn the parameters Θ by optimizing the following large-margin formulation: $\min_{\Theta\geq 0} \frac{1}{N}\sum_{n=1}^N \mathcal{L}_n(\Theta) + \frac{\lambda}{2}||\Theta||^2$, where $\mathcal{L}_n(\Theta)$ is the generalized hinge loss associated with training example n: $\mathcal{L}_n(\Theta) = \max_{\mathcal{Y}\subset\mathcal{V}^{(n)},|\mathcal{Y}|\leq k}(F(\mathcal{Y}, x^{(n)}, \Theta) + l_n(\mathcal{Y})) - F(\mathcal{Y}^{(n)}, x^{(n)}, \Theta)$. Here, $\mathcal{Y}^{(n)}$ is a human summary for the n^{th} ground set $\mathcal{V}^{(n)}$ (video, image collection, or text document), with corresponding features $x^{(n)}$. The specific objective functions and gradient computations in case of query-focused, privacy-preserving, and joint query-focused and privacy-preserving summarizations are presented in Appendix F. For generic summarization, we add the standard submodular functions modeling representation, diversity, coverage, *etc.* in the mixture. For query-focused summarization and privacy-preserving summarization, we instead use the MI and CG versions of the functions respectively as defined above[1]. Once the parameters are learned, we instantiate the mixture model with the parameters and maximize it to get the desired summaries.

3.3 Connections to Past Work

PRISM generalizes past work in both targeted learning and guided summarization. We summarize the connections below; for more details, see Appendix G.

Targeted Learning: A number of approaches like GLISTER (Killamsetty et al. 2020) and GRAD-MATCH (Killamsetty et al. 2021) can be considered with a validation set, and hence used in the targeted setting. Similarly, CRAIG (Mirzasoleiman, Bilmes, and Leskovec 2020) can be extended to consider a validation set. In Appendix G.1, we show that Algo. 1 in fact generalizes CRAIG (using FLQMI), GLISTER (using COM), and GRAD-MATCH (using GCMI + Diversity).

Guided Summarization: A number of past works for summarization have inadvertently used instances of PRISM. The query-DPP considered in (Sharghi, Gong, and Shah 2016; Sharghi, Laurel, and Gong 2017) is a special case of LOGDETMI. Similarly, the graph-cut based query-relevance term in (Vasudevan et al. 2017; Lin 2012), and in (Li, Li, and Li 2012) is actually GCMI. Furthermore, the joint diversity and query-relevance term in (Lin and Bilmes 2011) is an instance of COM (with the square-root as the concave function, see Appendix G.2).

4 Experiments and Results

4.1 Targeted Learning

In this section, we demonstrate the effectiveness of PRISM for targeted learning on the CIFAR-10 (Krizhevsky, Hinton et al. 2009), MNIST (LeCun et al. 1998), SVHN (Netzer et al. 2011), and P-MNIST (Pneumonia-MNIST) (Yang, Shi, and Ni 2021; Kermany et al. 2018) image classification datasets.

Custom dataset: To simulate a real-world setting, we randomly select some *target* classes and split the train set into *labeled*, *target*, and an *unlabeled* set such that (i) the *labeled* set has *class imbalance* and poorly represents the target classes, (ii) the *target* set has a small number of data points from the target classes, and (iii) the *unlabeled* set is a large set whose labels we do not use (resembling a large pool of unlabeled data in real-world scenarios). For CMI functions, we additionally use a *private set*, which has a small number of data points from the non-target classes. The performance is measured on the standard test set from the respective datasets. Let the set $\mathcal{C}$ consist of data points from the target classes and the set $\mathcal{D}$ consist of data points from the non-target classes. We create an initial labeled set $\mathcal{E}$, such that $|\mathcal{D}_{\mathcal{E}}| = \rho|\mathcal{C}_{\mathcal{E}}|$ and an unlabeled set which follows the same distribution $|\mathcal{D}_{\mathcal{U}}| = \rho|\mathcal{C}_{\mathcal{U}}|$, where ρ is the imbalance ratio. We use $\rho = 20$ and $|\mathcal{T}| = 10$ (total number of samples from target classes) for all experiments. For CIFAR-10,

[1]For the query-focused and the privacy-preserving case, CMI degenerates to MI and CG, respectively.

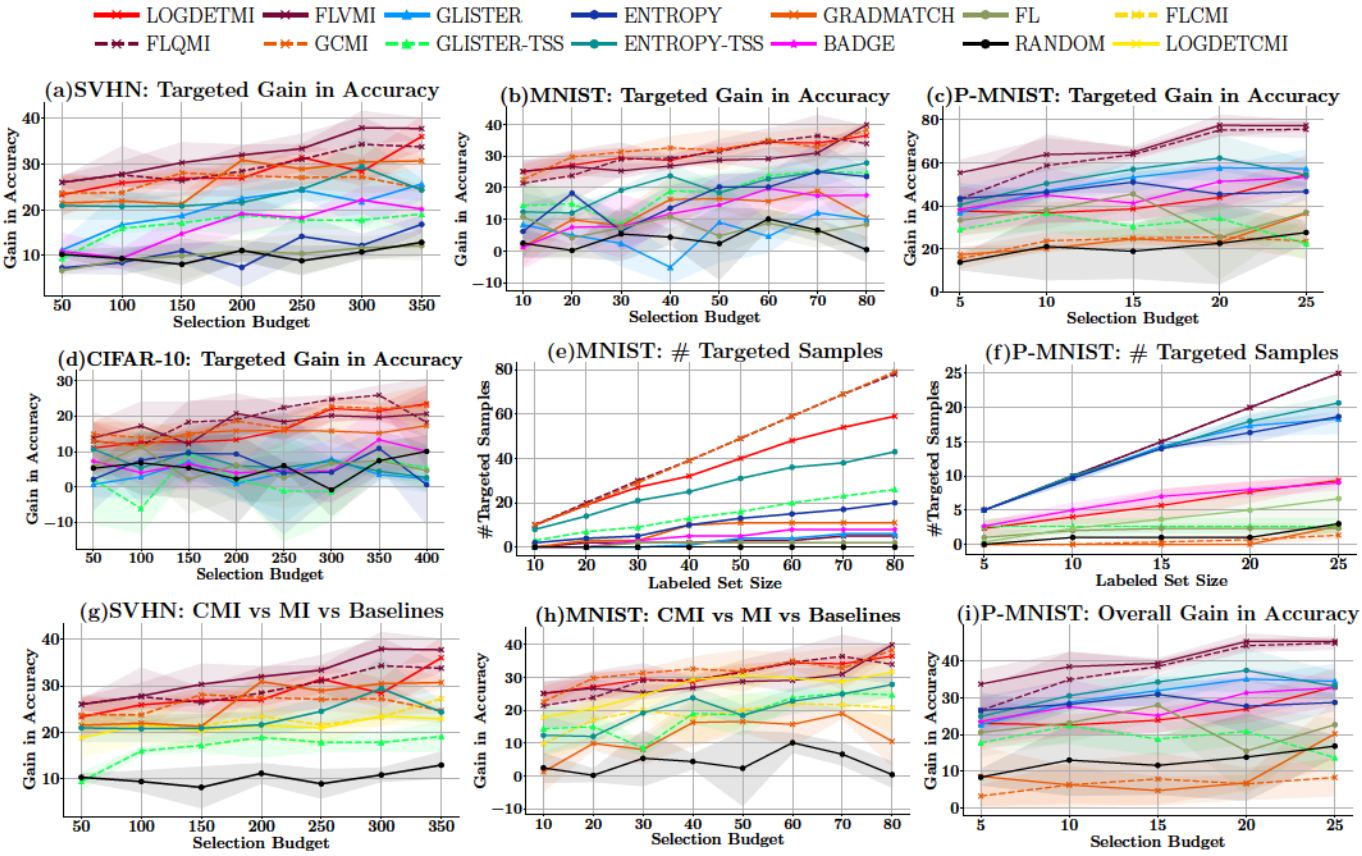

Figure 3: Targeted learning with PRISM on SVHN, MNIST, P-MNIST (Pneumonia-MNIST) and CIFAR-10. Plots (a-d) compare average gain in accuracy on targeted classes. Plots (e-f) compare the number of data points selected from the targeted classes. Plots (g-h) show an ablation study to compare the performance of MI and CMI functions. Plot (i) shows the overall gain in accuracy on P-MNIST. The MI and CMI functions (specifically FLQMI and FLVMI) obtain larger gains in targeted (plots a-d) and overall accuracy (plot i) than other baselines by selecting larger and more diverse examples from targeted classes (plots e-f).

MNIST and SVHN, we randomly select 2 classes as targets, while for the binary classification task in P-MNIST, we select the *pneumonia* class as the target. For MNIST and SVHN, $|\mathcal{C}_\mathcal{E}| + |\mathcal{D}_\mathcal{E}| = 1620$, $|\mathcal{C}_\mathcal{U}| + |\mathcal{D}_\mathcal{U}| = 24.3K$. For CIFAR-10, $|\mathcal{C}_\mathcal{E}| + |\mathcal{D}_\mathcal{E}| = 8400$, $|\mathcal{C}_\mathcal{U}| + |\mathcal{D}_\mathcal{U}| = 24.3K$. For P-MNIST, $|\mathcal{C}_\mathcal{E}| + |\mathcal{D}_\mathcal{E}| = 105$, $|\mathcal{C}_\mathcal{U}| + |\mathcal{D}_\mathcal{U}| = 1100$. These data splits were chosen to simulate low accuracy on target classes and at the same time to maintain the imbalance ratio in labeled and unlabeled datasets.

Baselines and Implementation details: We compare use of MI and CMI functions in Algo. 1 with other existing approaches. Specifically, for MI functions we use LOGDETMI, GCMI, FLVMI and, FLQMI. As baselines, we consider acquisition functions from the active learning literature; *viz.*, entropy sampling (ENTROPY), BADGE (Ash et al. 2020), and GLISTER-ACTIVE (Killamsetty et al. 2020). We run the active learning baselines only for one iteration to be consistent with our targeted learning setting (*i.e.*, we select from the unlabeled set only once). Since these active learning baselines do not explicitly have information of the target set, to further strengthen the comparison, we also compare against two variants that are target-aware. The first is 'targeted entropy sampling' (ENTROPY-TSS), where a product of the uncertainty and the similarity with the target is used to identify the subset, and the second is GLISTER for targeted subset selection (GLISTER-TSS), where the target set is used in the bi-level optimization. We also compare against GRADMATCH (Killamsetty et al. 2021), which mines for a subset such that the weighted difference in the gradients with the target set is minimized. Lastly, we also include vanilla FL and random sampling as baselines. For all datasets except MNIST, we train a ResNet-18 model (He et al. 2016). For MNIST, we train a LeNet model (LeCun et al. 1989). We use the cross-entropy loss and the SGD optimizer until training accuracy exceeds 99%. After augmenting the train set with the labeled version of the selected subset and re-training the model, we report the average gain in accuracy for the target classes and the overall gain in accuracy across all classes. The numbers reported are averaged over 10 runs of randomly picking any two classes for the target. We run Algo. 1 for different budgets and also study the effect of budget on the performance. We set the internal parameters to default values of 1. All experiments were run on an NVIDIA RTX 2080Ti GPU.

Results: We report the effect of budget on the average gain in accuracy of the target classes in Fig. 3(a-d). On all datasets, MI functions yield the best improvement in terms of accuracy on the target classes, *viz.*, $\approx$ 20-30% gain over the model's

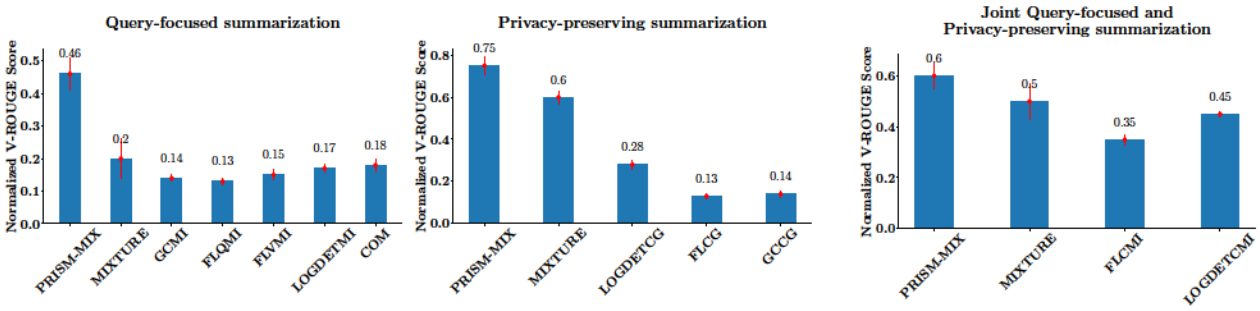

Figure 4: Guided summarization results on a real-world image collections dataset: because of the joint learning of the parameters, the proposed model (PRISM-MIX) outperforms others in all flavors of summarization.

performance before re-training with the added targeted subset. While this gain is $\approx 12\%$ higher than that of other methods, this also simultaneously improves the overall accuracy by $\approx$ 2-10% over other methods. Owing to their richer modeling, the MI functions consistently outperform all baselines across all budgets. This is also evident by the fact that MI functions select the most number of data points from the targeted classes (see Fig. 3(e-f)). Further, recall the discussion on the behavior of different MI functions in Section 2. As expected, FLVMI, FLQMI and LOGDETMI functions that model both query-relevance and diversity, perform better than a) functions which tend to prefer relevance (*viz.*, GCMI, ENTROPY-TSS) and b) functions which tend to prefer diversity/representation (*viz.*, BADGE and FL). Also, we observe that as the budget is increased, the MI functions outperform other methods by greater margins on the target class accuracy (Fig. 3). We run targeted learning for higher budgets on all datasets, and we observe that the MI functions achieve $20\times$ to $50\times$ labeling efficiency in obtaining the same accuracy on rare classes when compared to random and $2\times$ to $4\times$ compared to the best performing baseline (see Appendix H for more details). Additionally, we perform an ablation study to compare the performance of MI functions with the CMI functions and observe that they are at par with each other (see Fig. 3(g-i)). Finally, we do a pairwise t-test to compare the performance of all methods and observe that the MI functions (particularly FLVMI and FLQMI) statistically significantly outperform all baselines (see Appendix H). From a computational perspective, FLQMI and GCMI are the fastest in terms of running time and scalability and hence FLQMI is the preferred MI function given its scalability and consistent performance.

4.2 Guided Summarization

Dataset and Implementation Details: We use the image-collections dataset of (Tschiatschek et al. 2014). The dataset has 14 image collections with 100 images each and provides 50-250 human summaries per collection. We extend it by acquiring dense noun concept annotations (objects and scenes) for every image by pseudo-labelling using pre-trained off-the-shelf networks (Yolov3 pre-trained on OpenImagesv6 and ResNet50 pre-trained on Place365) followed by human correction. We designed query and private sets in a spirit similar to (Sharghi, Laurel, and Gong 2017) and acquired query-focused, privacy-preserving, and joint query-focused and privacy-preserving human summaries for every image collection. To instantiate the mixture model components, we extract concepts from images using the aforementioned pre-trained off-the-shelf networks and represent them, as well as the concept queries, by a $|\mathcal{C}|$-dimensional vector, where $\mathcal{C}$ is the universe of concepts. Our mixture model (PRISM-MIX) has four components which are the appropriate instantiations (MI/CG/CMI) of functions - GC, LogDet, FL and COM. The mixture weights as well as the internal parameters (λ, ν, η) are learned using the *train* set. Following (Tschiatschek et al. 2014), we perform leave-one-out cross validation and report average V-ROUGE across 14 runs. We also normalize V-ROUGE *s.t.* the human average is 1 and the random average is 0. We provide further details of the dataset and implementation in Appendix I.

Results: We present the guided summarization results in Fig. 4. As discussed in Section 3.3, the individual components of our mixture model have been used as models in previous works on document and video summarization. Hence, we compare the performance of PRISM-MIX against the performance of the individual components as our baselines. Also, to confirm the positive effect of jointly learning the parameters of PRISM along with the mixture weights, we compare PRISM-MIX against a mixture model (MIXTURE) of exactly the same components, but with only the model weights w being learned. Other internal parameters (λ, η, ν) are set to fixed values of 1. We observe that PRISM-MIX outperforms other techniques, including MIXTURE on all flavors of summarization (see Fig. 4). This is expected, as the joint learning of parameters offered by PRISM (Sec. 3.2) enables producing summaries that can better imitate the complexities of the ground-truth summaries.

5 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. We showed that PRISM has interesting connections to several past work, further reinforcing its utility. Through several experiments on targeted learning and guided summarization for diverse datasets, we empirically verified the superiority of PRISM over existing methods.

References

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *ICLR*.

Celis, L. E.; and Keswani, V. 2020. Implicit Diversity in Image Summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–28.

Chali, Y.; Tanvee, M.; and Nayeem, M. T. 2017. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 418–424.

Fujishige, S. 2005. *Submodular functions and optimization*. Elsevier.

Gupta, A.; and Levin, R. 2020. The Online Submodular Cover Problem. In *ACM-SIAM Symposium on Discrete Algorithms*.

Gygli, M.; Grabner, H.; and Gool, L. 2015. Video summarization by learning submodular mixtures of objectives. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3090–3098.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Iyer, R.; Khargoankar, N.; Bilmes, J.; and Asanani, H. 2021. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, 722–754. PMLR.

Ji, Z.; Xiong, K.; Pang, Y.; and Li, X. 2019. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jiang, P.; and Han, Y. 2019. Hierarchical Variational Network for User-Diversified & Query-Focused Video Summarization. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 202–206.

Kaushal, V.; Iyer, R.; Kothawade, S.; Mahadev, R.; Doctor, K.; and Ramakrishnan, G. 2019a. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1289–1299. IEEE.

Kaushal, V.; Iyer, R.; Kothawade, S.; Subramanian, S.; and Ramakrishnan, G. 2019b. A Framework Towards Domain Specific Video Summarization. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 666–675.

Kaushal, V.; Iyer, R. K.; Doctor, K.; Sahoo, A.; Dubal, P.; Kothawade, S.; Mahadev, R.; Dargan, K.; and Ramakrishnan, G. 2019c. Demystifying Multi-Faceted Video Summarization: Tradeoff Between Diversity, Representation, Coverage and Importance. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 452–461.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131.

Killamsetty, K.; Sivasubramanian, D.; Mirzasoleiman, B.; Ramakrishnan, G.; De, A.; and Iyer, R. 2021. GRAD-MATCH: A Gradient Matching Based Data Subset Selection for Efficient Learning. *In ICML*.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; and Iyer, R. 2020. GLISTER: Generalization based Data Subset Selection for Efficient and Robust Learning. *arXiv preprint arXiv:2012.10630*.

Kirchhoff, K.; and Bilmes, J. 2014. Submodularity for Data Selection in Machine Translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Krause, A.; and Golovin, D. 2014. Submodular function maximization.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, J.; Li, L.; and Li, T. 2012. Multi-document summarization via submodularity. *Applied Intelligence*, 37(3): 420–430.

Lin, H. 2012. *Submodularity in natural language processing: algorithms and applications*. Ph.D. thesis.

Lin, H.; and Bilmes, J. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 510–520.

Lin, H.; and Bilmes, J. A. 2012. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*.

Mirzasoleiman, B.; Badanidiyuru, A.; Karbasi, A.; Vondrák, J.; and Krause, A. 2015. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Mirzasoleiman, B.; Bilmes, J.; and Leskovec, J. 2020. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 6950–6960. PMLR.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming*, 14(1): 265–294.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Ozkose, Y. E.; Celikkale, B.; Erdem, E.; and Erdem, A. 2019. Diverse Neural Photo Album Summarization. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. IEEE.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Sharghi, A.; Gong, B.; and Shah, M. 2016. Query-focused extractive video summarization. In *European Conference on Computer Vision*, 3–19. Springer.

Sharghi, A.; Laurel, J. S.; and Gong, B. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4788–4797.

Singh, A.; Virmani, L.; and Subramanyam, A. 2019. Image Corpus Representative Summarization. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 21–29. IEEE.

Tschiatschek, S.; Iyer, R. K.; Wei, H.; and Bilmes, J. A. 2014. Learning mixtures of submodular functions for image collection summarization. In *Advances in neural information processing systems*, 1413–1421.

Vasudevan, A. B.; Gygli, M.; Volokitin, A.; and Van Gool, L. 2017. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, 582–590.

Wei, K.; Iyer, R.; and Bilmes, J. 2015. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, 1954–1963. PMLR.

Xiao, S.; Zhao, Z.; Zhang, Z.; Yan, X.; and Yang, M. 2020. Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. In *AAAI*, 12426–12433.

Yang, J.; Shi, R.; and Ni, B. 2021. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 191–195. IEEE.

Yao, J.-g.; Wan, X.; and Xiao, J. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2): 297–336.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 487–495.