RD-DPP: Rate-Distortion Theory Meets Determinantal Point Process to Diversify Learning Data Samples

Anonymous WACV Algorithms Track submission

Paper ID 712

Abstract

Selecting representative samples plays an indispensable role in many machine learning and computer vision applications under limited resources (e.g., limited communication bandwidth and computational power). Determinantal Point Process (DPP) is a widely used method for selecting the most diverse representative samples that can summarize a dataset. However, its adaptability to different tasks remains an open challenge, as it is challenging for DPP to perform task-specific tuning. In contrast, Rate-Distortion (RD) theory provides a way to measure task-specific diversity. However, optimizing RD for a data selection problem remains challenging because the quantity that needs to be optimized is the index set of the selected samples. To tackle these challenges, we first draw an inherent relationship between DPP and RD theory. Our theoretical derivation paves the way to take advantage of both RD and DPP for a task-specific data selection. To this end, we propose a novel method for task-specific data selection for multi-level classification tasks, named RD-DPP. Empirical studies on seven different datasets using five benchmark models demonstrate the effectiveness of the proposed RD-DPP method. Our method also outperforms recent strong competing methods, while exhibiting high generalizability to a variety of learning tasks [1].

1. Introduction

Even in the big-data era, selecting data samples is still a significant problem in resource-limited scenarios, where the computational resources or the transmission bandwidth are constrained. This matter is critical in a family of applications such as image processing and unmanned Aerial Systems (UAS), where data collection and transmitting capacity is highly constrained by limited power and networking resources. A higher data diversity, even in potentially unknown representation space, is known to boost the prediction power of Machine Learning (ML) algorithms. A powerful tool to enhance diversity is the Determinantal Point Process (DPP) [4, 7, 12, 23], which offers a formal approach to model diversity by quantifying dissimilarity among elements within a set, potentially in some latent feature space. It is widely used by the *machine learning* community in search engines, recommender systems [6], document summarization [30], and more recently in learning-based image processing [24] and regression models [13, 36]. A related concept is the Rate-Distortion (RD) theory commonly used by the *information theory* community to design and evaluate Source Codes (SC) for lossy data compression [9]. It characterizes the minimum compression rate for a tolerable distortion level based on the distribution geometry of data samples.

In this paper, we first reveal the inherent relationship between DPP and RD theory. The relation between RD and DPP comes from the fact that both methods are used to evaluate data diversity but from different perspectives. DPP evaluates the diversity by modeling the dissimilarity among samples in a set, while RD quantifies the minimum representation bits per sample (i.e. the compressibility of the samples) required for a given distribution to satisfy a certain distortion limit. Therefore, they are intrinsically related. This relationship has yet to receive the deserved attention from the research community. This study uses this relationship to design a new data selection policy for classification tasks.

Particularly, we realize that although there exists effective and approximately optimal DPP-based inference [6, 15, 17, 36], DPP is not task-oriented and considers merely the inherent diversity of data samples. Hence, data selected based on DPP may not necessarily yield the highest performance for different learning tasks. In contrast, authors in [37, 40, 41] find that RD-theory is a useful tool to measure the quality of representation for classification. However, maximizing RD-based measurement is challenging in

[1] The source code is available on https://anonymous.4open.science/r/RD-DPP-83DB

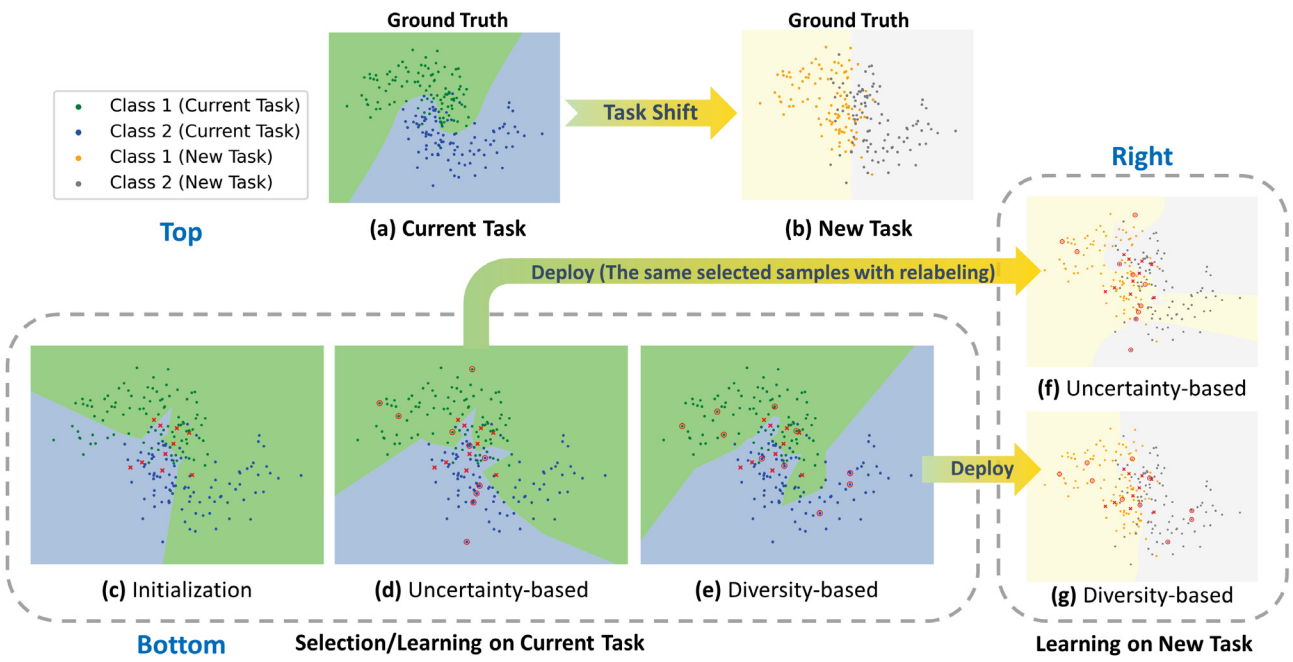

Figure 1. The benefit of using diversity-based methods in low-budget conditions; **Top:** ground truth decision boundary is shown for the *current task* (a) and *new task* (b). **Bottom:** (c-e): decision boundary learned using a kNN classifier for three scenarios using 10 initial random samples, marked as 'x', in (c); 10 initial samples + 10 uncertainty-based samples (selected based on (c), marked as 'o') in (d), and 10 initial samples + 10 diversity-based samples (selected based on (c), marked as 'o') in (e). The diversity-based method (e) is superior for mimicking distribution geometry than the uncertainty-based method. **Right:** compares the generalizability of different methods by applying the selected samples for the *current task* to the *new task*. The diversity-based method (e→g) that captures the overall geometry features is more generalizable than the uncertainty-based method (d→f), which excessively focuses on specific decision boundaries.

data selection since the variable in the optimization (i.e., the indices of selected samples) is discrete and unable to be optimized via the gradient-based method. Fortunately, our observed relationship provides the possibility to take mutual benefits between RD and DPP, and accordingly, we develop a novel algorithm called RD-DPP that facilitates sequential data selection. Specifically, we use class-conditional RD to measure the task-oriented semantic diversity (e.g., for a classification task) and perform Maximum a Posteriori (MAP) inference for DPP with RD-based quality-diversity kernel. The quality score of the kernel quantifies the added diversity of the sets with new samples with respect to the previously selected samples. After the semantic diversity is saturated, we use uncertainty methods to continually collect samples around the decision boundary if the transmission budget is available.

In summary, our contribution is two-fold: **(i)** It is the first work to reveal a concrete yet non-trivial relationship between Rate-Distortion theory and DPP under the mild assumption of Gaussianity; and **(ii)** We propose a novel data selection method for classification tasks by leveraging the relations between RD and DPP. The results in section 4 show that our method outperforms all alternative methods, including random selection, DPP-based methods, uncertainty-based methods, submodular mutual information methods, and density-based methods by a significant margin. Afterward, we demonstrate pursuing diversity is also beneficial for potential future tasks. The corresponding experiment is shown in Section 5. The comparison between our method and uncertainty-based methods, the most intuitive classification-oriented method, is exhibited in shown in Fig. 1.

2. Related Work

The data selection methods can be roughly divided into diversity-based methods and uncertainty-based methods. Along with DPP, there are several works that try to select the data based on measuring the diversity from different perspectives. For example, authors in [20, 32] based on density measurement and aim to select the samples approximately covering the entire distribution. Likewise, authors in [2, 19, 21] employ different submodular mutual information functions to measure the diversity. However, DPP and these methods are not task-oriented and hence challenged by the aforementioned issue that may not achieve optimal performance for different learning tasks.

An alternative approach to data selection is using

uncertainty-based methods by selecting data samples that are less consistent with the trained model based on metrics like *cross-entropy* and *margin* [8, 18, 29, 31]. An issue with this approach is its sensitivity to initial samples, causing poor early-stage performance until sufficient diverse samples are collected to establish reliable decision boundaries. Therefore, uncertainty-based methods are more advantageous when existing data is diverse enough.

A conceptual illustration of this phenomenon is presented in Fig. 1, which showcases a non-spherical dataset under budget limitations when the model is built with a few samples selected using the uncertainty-based method (d) and diversity-based method (e). Prioritizing diversity (Fig. 1 (e)) to approximate the population's distribution can lead to a more effective decision boundary compared to emphasizing the reduction of uncertainty of new samples in the close proximity of the decision boundary, especially when the initial model is not reliable. Additionally, Fig. 1 (f-g) shows pursuing diversity is also more beneficial for potential future tasks. We discuss this matter in Section 5.

3. Method

At first glance, DPP is a probabilistic model developed for various machine learning applications, which seems to diverge from the information-theoretic lens of RD. However, our in-depth analysis reveals that these two methods are inherently related and converge in the sample selection problem. To this end, we propose an RD-based DPP method for sample selection, termed RD-DPP. This section is organized as follows. To make this paper self-contained, we first give a brief introduction to RD (Section 3.1) and DPP (Section 3.2). Then, we introduce the derivation of the RD-DPP for sample selection (Section 3.3).

3.1. Rate-distortion Theory

An arbitrary real number (e.g., samples of continuous-valued signals) requires an infinite number of bits for lossless representation, which is impractical in most communication and storage systems. In practice, we usually settle with lossy compression that allows some representation errors. Specifically, given an arbitrary source X, we can use nR bits to encode a sequence of n samples X^n with $f_n(X^n)$ (using a codebook of size 2^{nR}) and then decode it with $\hat{X^n} = g_n(f_n(X^n))$. Here R denotes the coding rate. The reconstruction error for a sample sequence x^n is defined as $d(x^n, \hat{x}^n) := 1/n \sum_{i=1}^{n} d(x_i, \hat{x}_i)$ for some distance measure $d(\cdot, \cdot)$. A commonly used distortion metric is Mean Squared Errors (MSE) $\epsilon^2 := 1/n \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$ and distortion D is defined as $D := \mathbb{E}[d(X^n, \hat{X}^n)]$ [9]. In these notations, the uppercase letters are used for random variables, while lowercase letters denote their realizations. Rate distortion theory is used to quantify the minimum number of representation bits per sample R for a sequence with infinite length ($n \to \infty$) and distortion D. In our work, we use the estimated RD for a finite set of i.i.d. Gaussian distributed samples defined as follows.

Definition 1. Assume a finite dataset is represented by $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n] \in \mathbb{R}^{d \times n}$ (can be in some potentially learnable feature space) presenting n i.i.d. data points, each with d features, sampled from a zero-mean multivariate Gaussian distribution with covariance $\mathbf{\Sigma}$. The theoretical coding rate $R(\mathbf{Z}, \epsilon) := \frac{1}{2} \log \det \left(\frac{d}{\epsilon^2} \mathbf{\Sigma} \right)$ for a very small tolerable distortion ϵ^2 in squared error (SE) sense, can be approximately estimated as [27]

$$R(\mathbf{Z}, \epsilon) := \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{n\epsilon^2} \mathbf{Z}\mathbf{Z}^\top \right), \tag{1}$$

where the unit of $R(\mathbf{Z}, \epsilon)$ is bit/dimension for log base 2.

Remark 1. We use the term *approximate diversity* instead of *rate* when using metric $R(\mathbf{Z}, \epsilon)$ to emphasize that it is an approximate empirical measure computed for a finite set without using the parameters of its distribution.

For labeled data, each class can be compressed/encoded separately.

Definition 2. The coding rate of the sub-space for each class $R_i^c(\mathbf{Z}, \epsilon \mid C_i)$ is given by,

$$R_i^c(\mathbf{Z}, \epsilon \mid C_i) := \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{|C_i|\epsilon^2} \mathbf{Z}_{C_i} \mathbf{Z}_{C_i}^\top \right), \tag{2}$$

where C_i is the index set of class i, c_T is the number of classes, $\mathbf{Z}_{C_i}$ is a matrix using columns of $\mathbf{Z}$ indexed by C_i ($\mathbf{Z}[:, C_i]$), and $|C_i|$ is the cardinality of C_i.

3.2. Determinantal Point Processing (DPP)

Definition 3. DPP is a probability measure on all $2^{|\mathcal{A}|}$ subsets of $\mathcal{A}$, where $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. According to the definition of DPP [23], an arbitrary subset $A \subseteq \mathcal{A}$ drawn from $\mathcal{A}$ must satisfy,

$$\mathcal{P}(A) \propto \det(\mathbf{L}_A), \tag{3}$$

where $\mathcal{P}(A)$ denotes the probability of selecting subset A from the entire set $\mathcal{A}$. $\mathbf{L}$ is a positive semi-definite (PSD) *Gram Matrix* defined as $\mathbf{L} = \mathbf{Z}^\top \mathbf{Z}$ to measure pairwise similarity among points, and $\mathbf{L}_A$ is a submatrix of $\mathbf{L}$ with rows and columns indexed by set A.

We need normalization factor $\det(\mathbf{L} + \mathbf{I})$ when computing exact probabilities, because

$$\sum_{A \subseteq \mathcal{A}} \det(\mathbf{L}_A) = \det(\mathbf{L} + \mathbf{I}), \tag{4}$$

where A denotes a subset drawn from the entire set $\mathcal{A}$, for any $A \subseteq \mathcal{A}$. This identity has been proved in multiple materials, such as [23]. For data selection purposes, we often expect to ensure k samples with the largest diversity. This problem is known as Maximum a Posteriori (MAP) inference for DPP presented as,

$$\max_{A \subseteq \mathcal{Z}} \det(\mathbf{L}_A), \quad s.t. \ |A| = k, \ k \leq \text{rank}(\mathbf{L}). \tag{5}$$

It is an NP-hard problem, and the common solution is using *greedy search*, which has been developed in [6, 17]. A recent popular fast-known exact greedy approach was proposed in [6], and we denote it as $A^* = DPP_m(\mathbf{L}, k)$.

3.3. RD-DPP for Sample Selection

Here, we provide a theoretical relationship between RD and DPP as well as derive the proposed RD-DPP for sample selection. A summary of our bi-modal algorithm is presented in Algorithm 1.

Relation Between Rate-Distortion Theory and DPP. Rate-distortion Theory and DPP are inherently related. Consider $\alpha := \frac{d}{n\epsilon^2} > 0$ in Eq. (1). The value of the approximate diversity by RD theory (presented in Definition 1) can be described by the sum of the subset probabilities measured by DPP (presented in Eq. (4)), as follows

$$\begin{aligned} R(\mathbf{Z}, \epsilon) &\overset{(a)}{=} \frac{1}{2} \log \det \left(\mathbf{I} + \alpha \mathbf{Z}\mathbf{Z}^\top \right) \\ &\overset{(b)}{=} \frac{1}{2} \log \det \left(\mathbf{I} + \alpha \mathbf{Z}^\top \mathbf{Z} \right) \\ &\overset{(c)}{=} \frac{1}{2} \log \sum_{X \subseteq \mathcal{Z}} \det \left(\mathbf{L}_X \right), \end{aligned} \tag{6}$$

where $\mathbf{L} = \alpha \mathbf{Z}^\top \mathbf{Z}$ can be viewed as the L-ensemble kernel matrix of DPP, and $\mathcal{Z} = \{1, 2, \cdots, n\}$ denotes the index set of $\mathbf{Z}$. This relation states that the $R(\mathbf{Z}, \epsilon)$ can be described as the sum of point process measurements $\det(\mathbf{L}_X)$, which reveals a diverse set of samples should have high diversity for all of its possible subsets. This numerical equivalent can be proved by Sylvester's determinant identity [34], and we provide a simple proof in Appendix A.

DPP Approaches to Solve RD Problem. Based on the inherent relationship between the DPP and RP, we develop an RD-based quality function to measure individual rate gain as follows. Given a previously selected data set $\mathbf{Z} \in \mathbb{R}^{d \times n}$, how to search a new sample set $\mathbf{D} = \{\mathbf{z}_{d_1}, \mathbf{z}_{d_2}, \cdots \mathbf{z}_{d_k}\}$ with indices $\mathcal{D} = \{d_1, d_2, \cdots, d_k\} \subseteq \mathcal{B}$ such that the resulting diversity of $\mathbf{Z}^{\mathcal{D}+} := [\mathbf{Z}, \mathbf{z}_{d_1}, \mathbf{z}_{d_2}, \cdots \mathbf{z}_{d_k}] \in \mathbb{R}^{d \times (n+k)}$ (i.e. the diversity measured by Definition 1) is maximized, where $\mathcal{A}$, $\mathcal{Z} \subset \mathcal{A}$, and $\mathcal{B} = \mathcal{A} \setminus \mathcal{Z}$, respectively, denote the index set of entire data, previously selected samples, and candidate samples. Based on our conclusion in Eq. (6), we can compute the diversity after selecting the set $\mathcal{D}$ as

$$\begin{aligned} R(\mathbf{Z}^{\mathcal{D}+}, \epsilon) &= \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{(n+k)\epsilon^2} {\mathbf{Z}^{\mathcal{D}+}}^\top \mathbf{Z}^{\mathcal{D}+} \right) \\ &= \frac{1}{2} \log \sum_{X \subseteq \mathcal{Z}^{\mathcal{D}+}} \det \left(\widetilde{\mathbf{L}}_X \right), \end{aligned} \tag{7}$$

where $\widetilde{\mathbf{L}} = \frac{d}{(n+k)\epsilon^2} \mathbf{Z}^\top \mathbf{Z}$. $\widetilde{\mathbf{L}}_X$ is the submatrix of $\widetilde{\mathbf{L}}$ indexed by X, and $\mathcal{Z}^{\mathcal{D}+} = \mathcal{Z} \cup \mathcal{D}$ is the index set of $\mathbf{Z}^{\mathcal{D}+}$. Our goal here can be stated as

$$\arg\max_{\mathcal{D}} R(\mathbf{Z}^{\mathcal{D}+}, \epsilon), \quad s.t. \ |\mathcal{D}| = k. \tag{8}$$

Since in Eq. (7), the logarithm base is 2, we can obtain $2^{2R(\mathbf{Z}^{\mathcal{D}+}, \epsilon)} = \sum_{X \subseteq \mathcal{Z}^{\mathcal{D}+}} \det \left(\widetilde{\mathbf{L}}_X \right)$, which is the sum of DPP-based measure across the all subsets $X \subseteq \mathcal{Z}^{\mathcal{D}+}$. We can obtain the similar term $2^{2R(\mathbf{Z}^{b^i+}, \epsilon)}$ for each candidate $b^i \in \mathcal{B}$ as $\sum_X \det(\mathbf{L}_X)$. Here, $X \subseteq \mathcal{Z}^{b^i+} := \mathcal{Z} \cup \{b_i\}$, and we have $(n+k)\widetilde{\mathbf{L}}_X = (n+1)\mathbf{L}_X$. Noting that $R(\mathbf{Z}^{b^i+}, \epsilon)$ is the individual diversity gain for b^i that has the memory of the selected data set $\mathbf{Z}$ but has no knowledge about other candidates, which cannot facilitate diversity among candidates. By applying the set operation, we can translate the problem to one that considers both the individual rate gain of each candidate and the diversity among candidates. To this end, we develop the following approach to solve the optimization problem in Eq. (8). The diversity vector is the feature of each sample $\mathbf{z}_i, i \in \mathcal{B}$, and the quality score $\Phi()$ evaluates the individual rate gain from the perspective of RD theory. We can use MAP inference for DPP with a quality-diversity kernel $\mathbf{K}$ to solve the problem (i.e. $\arg\max_{\mathcal{D}} \mathbf{K}_{\mathcal{D}}$, where $\mathbf{K}_{\mathcal{D}}$ is $\mathbf{K}$'s rows and columns indexed by $\mathcal{D}$.) as follows,

$$\mathbf{K}_{i,j} = \Phi(\mathbf{Z}^{b^i+}, \epsilon)\Phi(\mathbf{Z}^{b^j+}, \epsilon)(\mathbf{L}_{\mathcal{B}})_{i,j}, \tag{9}$$

where $\mathbf{L}_{\mathcal{B}} = \mathbf{Z}_{\mathcal{B}}^\top \mathbf{Z}_{\mathcal{B}}$ is the gram matrix across all m candidate samples with index $\mathcal{B} = \{b^1, b^2, \cdots, b^m\} = \mathcal{A} \setminus \mathcal{Z}$, $\mathbf{Z}^{b^i+} := [\mathbf{Z}, \mathbf{z}_{b^i}]$. $\Phi(\mathbf{Z}^{b^i+})$ is the RD-based quality function to quantify the individual gain obtained by adding this sample to the known set.

Task-oriented RD-based Kernel. To enhance the quality of training for a specific learning task, such as the classification task, we develop a *semantic diversity* kernel instead of the original class-independent DPP diversity. First, we adopt the assumption by [37, 41]: i) The distribution of any high-dimensional data (Fig. 2(a)) is typically supported on a low-dimensional manifold (Fig. 2(b)). ii) A good data

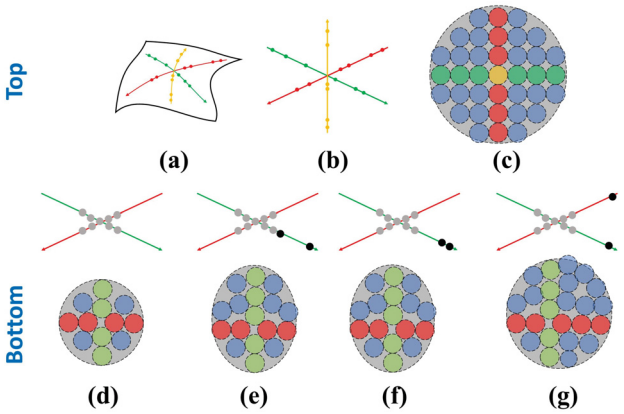

Figure 2. Visualizing the concept of task-oriented diversity. Top: Green, red, and yellow denote three different classes: (a) high-dimensional data, (b) corresponding low-dimensional manifold, and (c) *sphere packing*, where each sphere denotes a bit with ϵ^2 distortion tolerance. Here, yellow spheres are orthogonal on the plane. Bottom: Different scenarios for adding two new samples (black points) to the previously selected set of nine samples (gray points). (d) previously selected samples; (e-g) adding two samples based on the pure diversity of new samples (e), the individual marginal gain of the RD-based diversity (f). and the highest semantic diversity (g).

representation for a classification task should have *within-class diversity* and *between-class discrimination*. This is visualized by sphere packing in Fig. 2(c), where we expect the number of blue spheres to be large (maximize the between-class discrimination), and so is the number of the color spheres (except blue) to maximize the within-class diversity. Therefore, we can define semantic diversity for a set of samples $\mathbf{X}$ by class-conditional RD as

$$sdiv(\mathbf{X}) := R(\mathbf{X}, \epsilon) - \sum_{i=1}^{c_T} \frac{|C_i|}{n} R_i^c(\mathbf{X}, \epsilon \mid C_i) \geq 0. \quad (10)$$

where, again, C_i and c_T denote the index set of data in class i and total number of classes, respectively. Likewise, suppose a previously selected data set $\mathbf{Z}$. Like the Eqs. (8)-(10), to maximize the semantic diversity ($sdiv(\mathbf{Z}^{\mathcal{D}+}, \epsilon)$) by selecting an additional set $\mathcal{D}$ ($|\mathcal{D}| = k$), we can apply the RD-DPP relations and develop the RD-based quality function to evaluate the semantic diversity gain caused by selecting individual candidate $\mathbf{x}_i, i \in \mathcal{B}$ as,

$$\Phi(\mathbf{X}_{i+}, \epsilon) = sdiv(\mathbf{X}_{i+}), \quad (11)$$

where $\mathbf{X}_{i+} = [\mathbf{Z}, \mathbf{x}_i] \in \mathbb{R}^{d \times (n+1)}$. Then, similar to Eq. (9), the task-oriented DPP kernel $\mathbf{K}$ can be constructed based on Eqs. (9) and (11), as

$$\mathbf{K}_{i,j} = \Phi(\mathbf{X}_{i+}, \epsilon)\Phi(\mathbf{X}_{j+}, \epsilon)\langle \mathbf{x}_i, \mathbf{x}_j \rangle, \quad (12)$$

where $\langle \cdot\ ,\ \cdot \rangle$ denotes the inner product operation. Fig. 2(d)-(g) demonstrates different strategies to add two points to a known set of points for a labeled dataset, which indicates that we should take into account both individual diversity gain and the distance between the two candidates. A fast DPP MAP inference proposed in [6] can be used to search the k optimized candidates as $\arg\max_{\mathcal{D}} \mathbf{K}_{\mathcal{D}}$.

Bi-Modal Scheduling. As the number of selected samples increases, the increase in diversity caused by further sample selection will become increasingly gradual and eventually reach its upper bound asymptotically. Hence, at this time, selecting samples that are more uncertain to the learned decision boundary would yield greater advantages in refining the model further. Let us imagine people building a house. Typically, they first construct the framework (analogous to diversity-based selection) and then proceed to the interior finishing work (analogous to diversity-based selection).

To accommodate *Diversity Saturation* in the selection, we can set an empirical criterion to switch mode from RD-DPP diversity to uncertainty-based selection. To this end, we calculate the semantic diversity $sdiv(\mathbf{Z}^t)$ at the end of each round (for $t = k, 2k, 3k, \cdots$), and switch when we first observe $sdiv(\mathbf{Z}^t) - sdiv(\mathbf{Z}^{t-k}) < \phi_0$ meaning that the diversity improvement is less than a pre-defined threshold, ϕ_0. An ablation analysis for this bi-modal scheduling is presented in Section 4.

Algorithm 1 Bi-modal RD-DPP for Sample Selection

Input: Entire data with indices $\mathcal{A}$, Initial data $\mathbf{Z}_0$ with index set $\mathcal{Z}_0$, affordable transmission budget n_T samples, and the number of samples k selected in each round.

Output: The index set of selection $SelSet$.

Initialize: $SelSet \leftarrow \mathcal{Z}_0$, $\mathcal{B} \leftarrow \mathcal{A} \setminus \mathcal{Z}_0$, $\mathbf{Z} \leftarrow \mathbf{Z}_0$, and $t \leftarrow 0$, transitionFlag$\leftarrow False$.
while $t \leq n_T$ **do**
  $t \leftarrow t + k$ `#Mode one. DPP-based.`
  if $sdiv(\mathbf{Z}^t) - sdiv(\mathbf{Z}^{t-k}) > \phi_0$ **and Not** transitionFlag **then**
    Calculate the DPP kernel $\mathbf{K}$ for $\mathcal{B}$ by Eq. (12).
    $SelSet\_round \leftarrow DPP_m(\mathbf{K}, k)$.
  else
    transitionFlag$\leftarrow True$ `#To Mode two.`
    $SelSet\_round \leftarrow Uncertainty(\mathbf{x}_i, i \in \mathcal{B}, k)$
  end if
  $\mathcal{B} \leftarrow \mathcal{B}/SelSet\_round$
  $SelSet \leftarrow SelSet \cup SelSet\_round$
  for i **in** $SelSet\_round$ **do**
    $\mathbf{Z} \leftarrow [\mathbf{Z}, \mathbf{x}_i]$. `#Add one column to` $\mathbf{Z}$.
  end for
end while

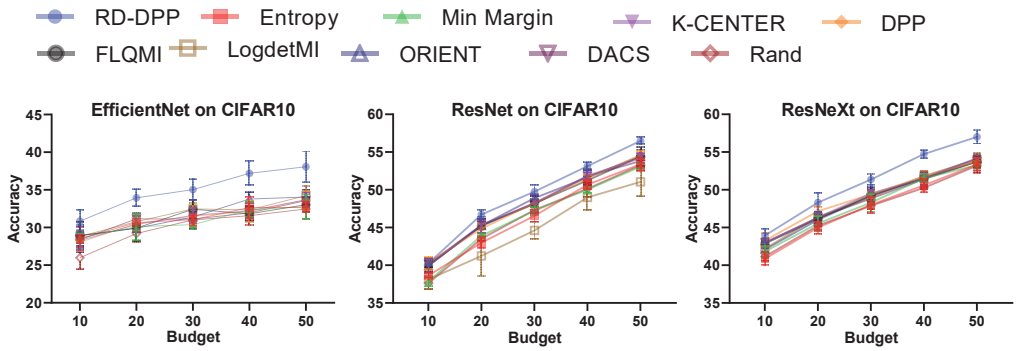

Figure 3. Performance of the classifier (%) with different network architectures for different selection methods applied to CIFAR10 dataset.

4. Experiment

Setup. As a proof of concept, we evaluate the proposed and alternative methods using seven datasets, including MNIST [11], FMNIST [38], CIFAR10 [22], SVHN [28], and three small datasets, Yeast [1], Cardiotocography [5], and Statlog (Landsat Satellite) [33]. It is worth mentioning that in practical communication systems, the data is transmitted in terms of packets that may contain more than one sample, and samples with a packet are very similar. To accommodate this consideration, we regulate the selection of packets (instead of samples) in our experiments. All the above derivations are valid with only one change that the sample's feature vectors $\mathbf{x}$ are replaced by the class-wise mean of feature vectors of all samples within the packet. In our experiment, we assume there exists a total of 100 packets for MNIST and FMNIST, and each packet contains 64 samples. Likewise, we constructed 100 packets for CIFAR10, and each packet had 200 samples. In each experiment, we use 5 randomly selected packets for initialization and then perform different strategies to select packets. Details of three small datasets and additional experiment setups are provided in Appendix C. The composition of some exemplary packets is shown in Appendix B.

To efficiently evaluate the effectiveness of our method with different models, we apply a simple CNN network on MNIST and FMNIST and use the feature from the layer before the classifier as the lower-dimensional representation of data samples (each image is mapped to a 288×1 vector). Likewise, we apply three different state-of-the-art architectures for CIFAR10: ResNet-18 [16], EfficientNet-B0 [35], and ResNeXt29 (2x64d) [39], respectively, representing each image as a vector of 320, 512, and 1024 elements. We apply a naive logistic regression for Yeast, Cardiotocography, and Statlog (Landsat Satellite).

Baselines. To prove the effectiveness, the proposed method should at least outperform the two most important baselines: i) **Random** selection and ii) **vanilla DPP**, which has been validated by [36] for this purpose. We also compare our method against multiple alternative selection policies. We include uncertainty-based methods [8, 18, 29] based on iii) **Cross-Entropy** and iv) **Min Margin**, respectively. For the sake of completeness, we also compare it against some diversity-based methods. These methods include v) **K-Center** selection [32], three submodular mutual information methods [21], vi) **FLQMI** [2], vii) **LogdetMI** and viii) **ORIENT** [19], viiii) and **DACS** [20], a density-based method.

Results. The main results (average of 10 runs) on four datasets are shown in Table 1, demonstrating our proposed method outperforms all other methods in most scenarios. The primary observation lies in the comparison with random selection. Our method yields a substantial accuracy improvement ranging a $3\% - 5\%$, $2\% - 12\%$, $2\% - 4\%$, and $3\% - 5\%$ accuracy gain at all budgets from 10 to 50 on four datasets, respectively, while random selection often beats other methods at some budgets. Our method can also obtain at least a 1% accuracy gain over other methods at transmission budgets up to 50. Additionally, when the budget is limited to 10 on the SVHN dataset, our approach achieves an impressive at least 7% increase in accuracy compared to all other methods. The results on CIFAR10 with different architectures are shown in Fig. 3 and Tbale 2, which exhibit a significant gain for our method over all other methods. For example, ResNet and ResNeXt, when using our selection strategy, obtain a 2%-3% gain over the entropy-based decision, min-margin decision, and random selection. EfficientNet obtains even a higher gain of 5% over the other methods at transmission budgets up to 50. We also observe that the diversity-based methods and ours outperform the uncertainty-based method at the beginning. However, when the transmission budget increases, these methods become even worse than the uncertainty-based methods, but our method consistently can outperform all of these methods. Please also refer to appendix C to see additional ex-

periments on raw samples for Yeast, Cardiotocography, and Statlog (Landsat Satellite).

Table 1. The comprehensive results on 4 datasets by different selection methods.

Dataset	Budget	10	20	30	40	50
MNIST	RD-DPP	**49.67±5.84**	**72.86±2.79**	**83.21±2.02**	**89.25±1.31**	**91.26±1.46**
	Cross-Entropy	22.42±3.70	45.43±3.41	70.23±4.08	87.15±2.06	90.23±1.35
	Min Margin	43.14±7.06	63.31±6.13	74.14±3.83	81.98±1.93	82.40±2.73
	K-CENTER	48.56±4.58	71.14±6.57	80.51±3.41	86.35±2.79	88.00±2.45
	vanilla DPP	49.47±3.58	69.22±3.61	80.30±5.78	85.88±4.01	89.43±2.04
	FLQMI	34.44±7.55	48.65±6.76	61.89±7.91	67.84±4.38	77.61±2.37
	LogdetMI	46.59±6.86	60.23±3.85	68.19±3.77	71.38±2.85	79.47±2.45
	ORIENT	21.84±3.51	29.38±2.64	46.48±5.77	56.09±4.65	69.59±4.81
	DACS	22.78±5.49	48.75±8.68	75.76±5.28	84.14±3.04	90.40±1.27
	Rand	44.12±5.61	64.50±3.37	78.58±4.59	84.29±3.23	87.91±1.89
FNMIST	RD-DPP	44.36±3.00	**50.75±2.78**	**54.75±2.10**	55.87±1.61	**59.35±1.08**
	Cross-Entropy	27.32±5.52	43.99±6.32	51.64±4.32	55.65±3.47	57.43±2.45
	Min Margin	30.39±7.70	39.97±3.65	45.14±6.48	47.07±5.53	51.12±2.62
	K-CENTER	43.08±5.45	49.43±1.84	54.20±3.54	55.05±3.51	58.35±1.88
	vanilla DPP	44.12±3.33	49.29±3.11	54.34±2.77	55.71±3.07	58.30±1.00
	FLQMI	44.16±4.90	50.03±3.16	53.91±3.28	55.80±3.04	56.38±3.38
	LogdetMI	44.38±3.93	50.20±3.89	53.68±4.39	55.62±4.02	57.88±3.50
	ORIENT	**45.04±6.76**	50.53±3.74	53.52±3.77	**56.98±3.63**	57.65±1.52
	DACS	44.20±2.24	49.03±5.10	51.66±3.39	55.78±4.28	56.45±2.36
	Rand	32.10±4.30	45.29±6.22	52.46±4.98	53.85±3.11	56.30±3.41
CIFAR10	RD-DPP	**43.94±0.88**	**48.31±1.3**	**51.36±0.72**	**54.74±0.51**	**57.02±0.89**
	Cross-Entropy	40.86±0.86	44.98±0.78	48.0±1.08	50.6±0.85	53.36±0.81
	Min Margin	41.83±0.76	45.54±0.96	48.3±1.18	51.5±0.58	53.28±0.6
	K-CENTER	42.98±0.67	46.32±0.63	49.52±0.52	51.5±0.82	54.2±0.31
	vanilla DPP	43.29±0.93	47.16±0.6	49.4±0.51	51.76±0.74	53.79±0.38
	FLQMI	42.87±0.53	46.16±0.47	49.28±0.88	51.54±0.86	53.73±0.68
	LogdetMI	42.19±1.29	46.40±0.55	49.01±0.79	51.85±0.60	54.00±0.83
	ORIENT	42.10±1.01	46.02±0.87	49.20±1.20	51.57±0.91	54.12±0.42
	DACS	43.02±0.35	46.31±0.74	48.84±0.29	51.43±0.75	53.74±0.88
	Rand	41.16±0.62	45.24±0.61	47.88±0.43	50.26±0.12	53.2±0.96
SVHN	RD-DPP	**47.69±2.48**	**68.68±1.76**	**71.56±1.07**	**74.92±0.53**	**77.14±0.35**
	Cross-Entropy	37.38±1.10	53.59±2.54	67.58±2.10	72.25±1.89	74.80±1.01
	Min Margin	37.35±2.03	53.98±1.30	66.10±1.35	71.49±1.03	74.48±1.26
	K-CENTER	37.50±2.19	52.97±2.17	65.89±2.15	71.28±1.63	73.96±1.12
	vanilla DPP	37.42±2.01	54.21±2.19	65.12±1.04	71.59±1.74	75.12±1.01
	FLQMI	37.16±1.77	53.67±1.97	65.80±2.01	72.13±0.87	75.01±0.64
	LogdetMI	40.19±1.92	55.12±2.12	67.90±2.21	71.88±0.81	75.02±0.98
	ORIENT	38.19 ±2.30	51.37±2.76	69.89 ±1.98	70.45±2.91	76.02 ±0.65
	DACS	37.28±2.30	53.19±1.30	65.79±2.28	72.64±1.19	74.12±0.94
	Rand	38.20±1.66	54.49±2.69	65.96±1.33	70.85±1.61	74.65±1.02

Table 2. Comparison of the classification accuracy by using different architecture on CIFAR10.

Network	Method	10	20	30	40	50
Efficient	RD-DPP	**30.85±1.5**	**33.95±1.13**	**35.03±1.42**	**37.19±1.61**	**38.07±2.02**
	Cross-Entropy	28.5±0.58	30.65±0.42	31.03±0.65	32.65±1.35	32.67±0.68
	Min Margin	28.6±0.67	30.12±1.74	30.38±0.65	32.35±1.01	32.79±1.61
	K-CENTER	28.26±1.28	30.41±1.07	31.62±0.98	32.24±0.99	33.68±0.95
	vanilla DPP	28.05±1.19	30.53±0.95	31.52±1.43	32.12±1.36	33.69±1.2
	FLQMI	28.95±1.79	29.93±1.67	32.43±1.24	31.83±0.96	33.03±0.96
	LogdetMI	28.83±0.86	30.78±0.42	32.47±0.82	32.35±0.98	34.24±1.33
	ORIENT	28.89±1.40	29.91±0.52	31.39±1.55	33.78±0.90	34.05±0.89
	DACS	28.36±1.67	31.15±0.78	31.11±0.97	31.95±0.93	33.58±0.55
	Rand	25.98±1.47	29.2±1.07	31.01±0.89	31.58±1.23	32.47±1.34
ResNet	RD-DPP	40.26±0.38	**46.71±0.62**	**49.75±0.93**	**53.13±0.51**	**56.49±0.49**
	Cross-Entropy	38.63±0.28	42.99±0.76	46.55±0.74	50.67±0.56	53.38±0.82
	Min Margin	37.89±0.66	43.91±0.22	47.24±0.55	50.03±0.4	53.08±0.42
	K-CENTER	39.92±0.95	45.33±1.07	48.26±1.22	51.87±0.46	54.19±0.66
	vanilla DPP	**40.36±0.8**	44.84±0.62	48.25±0.6	51.53±0.41	54.6±0.52
	FLQMI	40.01±0.92	45.20±0.89	48.22±0.50	51.24±0.55	54.42±1.22
	LogdetMI	38.12±1.27	41.24±2.63	44.61±1.17	48.99±1.71	51.05±1.89
	ORIENT	39.95±0.86	45.25±0.93	48.91±0.46	51.63±1.01	54.46±0.88
	DACS	40.15±0.38	45.32±0.56	48.11±1.10	51.81±0.92	53.83±0.68
	Rand	37.64±0.75	43.5±0.29	47.32±0.39	50.13±1.09	53.27±0.64
ResNext	RD-DPP	**43.94±0.88**	**48.31±1.3**	**51.36±0.72**	**54.74±0.51**	**57.02±0.89**
	Cross-Entropy	40.86±0.86	44.98±0.78	48.0±1.08	50.6±0.85	53.36±0.81
	Min Margin	41.83±0.76	45.54±0.96	48.3±1.18	51.5±0.58	53.28±0.6
	K-CENTER	42.98±0.67	46.32±0.63	49.52±0.52	51.5±0.82	54.2±0.31
	vanilla DPP	43.29±0.93	47.16±0.6	49.4±0.51	51.76±0.74	53.79±0.38
	FLQMI	42.87±0.53	46.16±0.47	49.28±0.88	51.54±0.86	53.73±0.68
	LogdetMI	42.19±1.29	46.40±0.55	49.01±0.79	51.85±0.60	54.00±0.83
	ORIENT	42.10±1.01	46.02±0.87	49.20±1.20	51.57±0.91	54.12±0.42
	DACS	43.02±0.35	46.31±0.74	48.84±0.29	51.43±0.75	53.74±0.88
	Rand	41.16±0.62	45.24±0.61	47.88±0.43	50.26±0.12	53.2±0.96

For further verifying the effectiveness of the proposed method by evaluating the average performance rank, we construct a critical difference diagram based on the Wilcoxon signed-rank test to detect pairwise significance with $\alpha = 0.05$ [10]. As shown in Fig. 4, the result demonstrates our method achieves a 1.15 average rank and confirms the statistical superiority over all other methods.

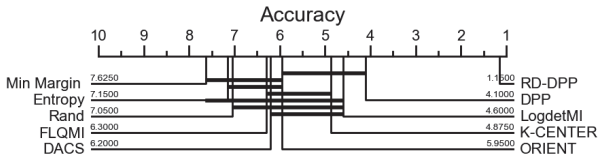

Figure 4. Critical difference diagram with $\alpha = 0.05$. There is no statistical difference between methods connected by a bolded line.

Ablation Analysis. We conduct an ablation study to demonstrate the effectiveness of the bi-modal scheduling. Two baselines are considered: i) RD-DPP (only diversity), which focuses solely on diversity without transitioning to uncertainty-based methods after diversity reaches the saturation point, ii) *Marginal Rate Gain*, which selects the k top candidate samples merely based on their individual semantic gains (i.e. k largest $\Phi(\mathbf{X}_{i+}, \epsilon)$ defined by Eq. (11)) ignoring the within-diversity of the candidate samples. The result in Table 3 shows that before the transition point, our bi-modal RD-DPP method outperforms the *Marginal Rate Gain* with 8%-10% and 5%-16% accuracy improvement on MNIST and FMNIST, respectively. Our method is equivalent to RD-DPP (only diversity) before the transition point as expected. After the phase transition point (i.e. the point between 20-30 and 40-50 for MNIST and FMNIST, respectively), RD-DPP (bi-modal) consistently achieves around 3% accuracy gain over the other two baseline methods.

Table 3. The ablation analysis on MNIST and FMNIST, respectively. Here, ✗ denotes there is no *Phase Transition*, while ✓ denotes that *Phase Transition* has occurred in the RD-DPP (Bi-modal).

Dataset	Budget	10	20	30	40	50	60
MNIST	Phase Trans?	✗	✗	✓	✓	✓	✓
	RD-DPP (Bi-modal)	**49.67**	**72.86**	**83.21**	**89.25**	**91.26**	**92.36**
	RD-DPP (Only Diversity)	-	-	80.92	84.72	87.29	90.11
	Marginal Rate Gain	41.99	63.13	72.3	83.03	84.71	86.74
FMNIST	Phase Trans?	✗	✗	✗	✗	✓	✓
	RD-DPP (Bi-modal)	**44.36**	**50.75**	**54.75**	**55.87**	**59.35**	**63.45**
	RD-DPP (Only Diversity)	-	-	-	-	56.36	57.58
	Marginal Rate Gain	28.12	42.94	48.64	54.81	57.86	60.56

Complexity Analysis. Our main overhead is to compute the semantic quality score (Eq. (11)). For each candidate i, the complexity of the term $R(\mathbf{X}_{i+}, \epsilon) = \log\det\left(\mathbf{I} + \alpha \mathbf{X}_{i+}\mathbf{X}_{i+}^{\top}\right)$ is only $\mathcal{O}(td\min(t,d))$ (i.e. the

complexity of the SVD decomposition of $\mathbf{X}_{i+}$). Therefore, we need operations in the order of $\mathcal{O}(mtd\min(t,d))$ to compute the semantic quality score of all candidates. Then, constructing the kernel presented in Eq. (12) requires a $\mathcal{O}\big((dc_T)m^2\big) = \mathcal{O}\big(dm^2\big)$ complexity for a small number of cluster/class labels c_T. The remaining complexity is the same as the greedy search method [6], which requires a $\mathcal{O}(k^2m)$ complexity to return k select samples. Thus, the overall complexity in each round is $\mathcal{O}(mtd\min(t,d) + d^2m + k^2m) \approx \mathcal{O}(mtd\min(t,d))$. In our work, we use bootstrapping to accelerate the approximation.

5. Discussion: Learning for Future Tasks

In this section, we highlight the broader advantages of diversity-based selection methods beyond their immediate benefits for current tasks by effectively preserving representative information. In contrast, uncertainty-based approaches, which are primarily designed to enhance the current model, lack this capacity. Here, we consider two scenarios:

Robustness for Label-shift Generalization. In this scenario, the same set of selected data samples are used for different tasks. We perform our experiment on the Large-scale CelebFaces Attributes (CelebA) Dataset [25], where each attribute can be used as the target label to perform a binary classification task. We select the samples for the `Smiling` classification task, then train two new classifiers for `Blond_Hair` and `High_Cheekbones` target labels using the same selected samples. The results after running 20 times are reported in Table 4 for 20, 40, and 60 selected packets out of 100 packets, where each packet contains 2 samples. The results show that our approach preserves diversity, confirming its potential to benefit various related tasks, especially under low-budget conditions.

Table 4. Generalizability of our RD-DPP and Uncertainty-based Decision (**Cross-Entropy**) methods are assessed by selecting samples for the original classification task (`Smiling`) and using them for two new classification tasks (`Blond_Hair` and `High_Cheekbones`). Methods are compared in F1 Score and Classification Accuracy.

Task	Budget	20		40		60	
	Method	F1	ACC	F1	ACC	F1	ACC
Smiling	RD-DPP	**63.71**	**63.97**	**70.04**	**70.24**	**72.45**	**72.79**
	Uncertainty Dec.	35.64	51.51	54.29	60.38	65.75	67.58
	Δ	+28.07	+12.46	+15.75	+9.86	+6.7	+5.21
Blond_Hair	RD-DPP	**64.48**	**89.5**	**66.92**	**90.01**	**70.41**	90.63
	Uncertainty Dec.	48.26	88.69	56.45	89.58	68.83	**90.87**
	Δ	+16.22	+0.81	+10.47	+0.43	+1.58	-0.24
High_Cheekbones	RD-DPP	**61.45**	**62.68**	**67.4**	**68.19**	**69.17**	**70.36**
	Uncertainty Dec.	40.47	54.32	53.88	60.82	60.84	65.11
	Δ	+20.98	+8.36	+13.52	+7.37	+8.33	+5.25

Table 5. The ability (classification accuracy) to resist negative interference on different tasks.

Task	Budgets	10	30	50
Rotated MNIST	RD-DPP	**53.30**	**57.27**	**69.78**
	Uncertainty Dec.	43.09	48.01	66.25
	Δ	+10.21	+9.26	+3.53
MNIST Fellowship	RD-DPP	**32.45**	**55.45**	**60.88**
	Uncertainty Dec.	17.05	38.29	56.27
	Δ	+15.4	+17.16	+4.61

Robustness for Domain-shift Interference. Real-world applications often involve data from different sources aiming at similar tasks (domain shift), such as classifying vehicles in urban and rural areas. In such cases, training one model for all tasks (i.e. multi-task learning) is not as effective as task-specific models due to the inherent variability among tasks. To alleviate this issue, it is advantageous to select samples that preserve task-specific information when switching between different domains under resource constraints [3]. We claim that our RD-DPP provides such capability. To this end, we construct two popular synthesis tasks, which are Rotated MNIST [26] and MNIST Fellowship [14]. For more details about the setup, please refer to Appendix B. The experimental results of the model training on the mixed data and inference on the original task are summarized in Table 5. The results clearly indicate that the proposed RD-DPP not only outperforms uncertainty-based decisions in addressing the original problem but also demonstrates the capability to reduce inter-domain interference in multi-task learning.

6. Conclusion

Our study reveals an inherent relationship between the RD and DPP when it comes to selecting diverse training samples to boost the performance of machine learning algorithms. This relationship is used to design a new measure of diversity for data that facilitates sequential DPP inference. We also propose bi-modal scheduling that switches between the DPP-based and uncertainty-based data selection modes to accommodate different transmission budget constraints better than all alternative selection methods. We showed that our approach can be applied to both raw data and data representation in low-dimensional latent spaces. The intensive experiment results using 7 different datasets and five different ML/DL models consistently show that our method outperforms pure uncertainty-based, pure diversity-based (including pure DPP-based), and random selection methods. Finally, we observed that the samples selected by our method are more beneficial (compared to other selection methods) for potential future tasks, such as label-shift tasks and domain-shift tasks.

References

[1] Yeast. UCI Machine Learning Repository, 1996. DOI: 10.24432/C5KG68. 6

[2] Nathan Beck, Truong Pham, and Rishabh Iyer. Theoretical analysis of submodular information measures for targeted data subset selection. *arXiv preprint arXiv:2402.13454*, 2024. 2, 6

[3] Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems*, 33:14879–14890, 2020. 8

[4] Daniele Calandriello, Michal Derezinski, and Michal Valko. Sampling from a k-dpp without looking at all items. *Advances in Neural Information Processing Systems*, 33:6889–6899, 2020. 1

[5] J. Campos, D. & Bernardes. Cardiotocography. UCI Machine Learning Repository, 2010. DOI: 10.24432/C51S4N. 6

[6] Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 4, 5, 8

[7] Xiwen Chen, Huayu Li, Rahul Amin, and Abolfazl Razi. Learning on bandwidth constrained multi-source data with mimo-inspired dpp map inference. *IEEE Transactions on Machine Learning in Communications and Networking*, 2024. 1

[8] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020. 3, 6

[9] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 1, 3

[10] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006. 7

[11] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 6

[12] Michał Dereziński. Fast determinantal point processes via distortion-free intermediate sampling. In *Conference on Learning Theory*, pages 1029–1049. PMLR, 2019. 1

[13] Michał Derezinski and Michael W Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1):34–45, 2021. 1

[14] Arthur Douillard and Timothée Lesort. Continuum: Simple management of complex continual learning scenarios. *arXiv preprint arXiv:2102.06253*, 2021. 8

[15] Julia Grosse, Rahel Fischer, Roman Garnett, and Philipp Hennig. A greedy approximation for k-determinantal point processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3052–3060. PMLR, 2024. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Shinichi Hemmi, Taihei Oki, Shinsaku Sakaue, Kaito Fujii, and Satoru Iwata. Lazy and fast greedy map inference for determinantal point process. *Advances in Neural Information Processing Systems*, 35:2776–2789, 2022. 1, 4

[18] Heinrich Jiang and Maya Gupta. Minimum-margin active learning. *arXiv preprint arXiv:1906.00025*, 2019. 3, 6

[19] Athresh Karanam, Krishnateja Killamsetty, Harsha Kokel, and Rishabh Iyer. Orient: Submodular mutual information measures for data subset selection under distribution shift. *Advances in neural information processing systems*, 35:31796–31808, 2022. 2, 6

[20] Yeachan Kim and Bonggun Shin. In defense of core-set: A density-aware core-set selection for active learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 804–812, 2022. 2, 6

[21] Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Submodular mutual information for targeted data subset selection. *ICLR 2021 From Shallow to Deep: Overcoming Limited and Adverse Data Workshop*, 2021. 2, 6

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[23] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. 1, 3, 4

[24] Claire Launay, Agnès Desolneux, and Bruno Galerne. Determinantal point processes for image processing. *SIAM Journal on Imaging Sciences*, 14(1):304–348, 2021. 1

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 8

[26] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 8

[27] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1546–1562, 2007. 3

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[29] Dongmin Park, Dimitris Papailiopoulos, and Kangwook Lee. Active learning is a strong baseline for data subset selection. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022. 3, 6

[30] Laura Perez-Beltrachini and Mirella Lapata. Multi-document summarization with determinantal point process attention. *Journal of Artificial Intelligence Research*, 71:371–399, 2021. 1

[31] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001. 3

[32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 2, 6

[33] Ashwin Srinivasan. Statlog (Landsat Satellite). UCI Machine Learning Repository, 1993. DOI: 10.24432/C55887. 6

[34] James Joseph Sylvester. Xxxvii. on the relation between the minor determinants of linearly equivalent quadratic functions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1(4):295–305, 1851. 4

[35] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[36] Nicolas Tremblay, Simon Barthelmé, and Pierre-Olivier Amblard. Determinantal point processes for coresets. *J. Mach. Learn. Res.*, 20:168–1, 2019. 1, 6

[37] Peng Wang, Huikang Liu, Druv Pai, Yaodong Yu, Zhihui Zhu, Qing Qu, and Yi Ma. A global geometric analysis of maximal coding rate reduction. In *Forty-first International Conference on Machine Learning*, 2024. 1, 4

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 6

[39] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[40] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *Advances in Neural Information Processing Systems*, 36:9422–9457, 2023. 1

[41] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020. 1, 4