BIG DATA MINING AND ANALYTICS
DOI: 10.26599/BDMA.2025.9020011
ISSN 2096-0654 0?/?? pp???–???
Volume 1, Number 1, January 2018

SPOT: An Active Learning Algorithm for Efficient Deep Neural Network Training

Luyang Fang, Cheng Meng, Lin Zhao, Tao Wang, Tianming Liu, Wenxuan Zhong*, and Ping Ma*

Abstract: Recent advancements in deep neural networks heavily rely on large-scale labeled datasets. However, acquiring these datasets can be challenging due to annotation constraints. Active learning offers a promising solution to this problem by selectively labeling a small, strategically chosen subset of the unlabeled dataset. However, current active learning methods struggle with data that is unevenly distributed, which leads to the selection of subsets that fail to represent the entire dataset. To overcome this challenge, we introduce a novel active learning algorithm that integrates space-filling (SP) designs with the optimal transport (OT) technique (SPOT). SPOT utilizes optimal transport to effectively manage data from complex manifolds by mapping it to a uniformly distributed hypercube. Additionally, the space-filling design facilitates a faster asymptotic convergence rate, ensuring that the selected subset encompasses the entire dataset more effectively than other sampling methods, such as random sampling. Our extensive experiments across various image datasets and models demonstrate the superiority of SPOT over existing baselines.

Key words: Active Learning; Optimal Transport; Space-Filling; Sampling; Deep Learning

1 Introduction

Deep neural networks (DNNs) have achieved significant advancements in various domains, including image recognition and natural language processing [1, 2, 3, 4, 5, 6]. The training of these large-scale DNNs typically requires extensive labeled datasets. For example, the optimization of the Vision Transformer relies on the JFT-300M dataset, which contains 303 million labeled samples [2]. However, acquiring annotations for such extensive training sets presents challenges due to cost, privacy, and the need for specialized expertise [7, 8, 9]. Active learning (AL) has recently emerged as a promising strategy to address these challenges by efficiently selecting a subset from the unlabeled pool for annotation, thereby optimizing the construction of training datasets [10, 11, 12, 13]. Unlike random sampling, which regards all data points as equally important, AL assumes that certain data points within the unlabeled pool are more critical for model optimization. The goal is to develop a learning algorithm that can identify and select these pivotal data points for annotation.

Current AL strategies for DNNs fall into two primary categories: uncertainty-based and diversity-based methods. Uncertainty-based methods focus on querying data points with high uncertainty, yet

• Luyang Fang, Tao Wang, Wenxuan Zhong, and Ping Ma are with the Department of Statistics, University of Georgia, Athens and 30602, USA. E-mail: luyang.fang@uga.edu, tw95546@uga.edu, wenxuan@uga.edu, pingma@uga.edu
• Cheng Meng is with the Institute of Statistics and Big Data, Renmin University of China, Beijing and 100080, China. E-mail: chengmeng@ruc.edu.cn
• Lin Zhao and Tianming Liu are with the Department of Computer Science, University of Georgia, Athens and 30602, USA. E-mail: lin.zhao@uga.edu, tliu@uga.edu
* To whom correspondence should be addressed. Email: {wenxuan, pingma}@uga.edu
Manuscript received: year-month-day; accepted: year-month-day

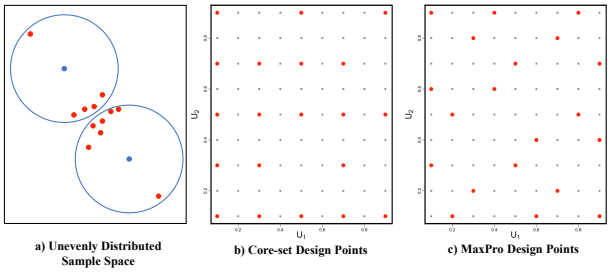

Fig. 1 a) Unevenly distributed sample space. Distance-based methods tend to select the points from the sparse areas (e.g., blue points) to cover the entire space, which ignores a lot of information (red points). b) Data points with core-set design. When projected to the U_1 dimension, 20 points are collapsed into only 5 points because of overlap. c) Data points with MaxPro design. In total 9 points are kept after the projection.

they risk selecting similar or duplicate samples. Conversely, diversity-based methods aim to encompass a comprehensive range of the sample space by selecting data points that maximize diversity based on their distances. A notable approach within this category is to select a subset from a core-set perspective [14, 15], which aims to represent the distribution of the entire dataset effectively [16, 17, 18]. For instance, Sener and Savarese [16] addresses the core-set selection challenge by formulating it as a minimax-based k-center problem [19]. The goal here is to determine k center points that cover the entire space by minimizing the maximum distance between any data point and its nearest center.

However, current core-set-based methods exhibit limitations in dealing with the data points that are unevenly distributed on the sample space, primarily because these methods do not estimate or account for distribution density. For example, core-set-based methods tend to select subsets that overly represent sparse areas in order to cover the entire sample space, consequently overlooking substantial information. As shown in Fig. 1 (a), points selected by the distance-based methods, represented in blue, result in a distorted representation of the original dataset. This can lead to subsets that do not accurately reflect the original dataset. Furthermore, minimax or maximin [20] distance designs are ineffective in projecting the selected design points onto subspaces. As illustrated in Fig. 1 (b), the representative data points from the original high-dimensional space tend to cluster and overlap when projected onto subspaces, leading to inefficient resource allocation at the subspace level. Given the principle of effect sparsity [21], which suggests that only a few dimensions in the data are statistically significant, it is crucial to project data accurately onto subspaces defined by these key dimensions. However, since the significant factors are not known in advance, ensuring an effective projection across all potential subspaces.

To address the aforementioned limitations, we introduce a novel diversity-based AL algorithm, named SPOT, which combines space-filling (SP) designs with optimal transport (OT) techniques. Optimal transport techniques [22, 23, 24, 25] efficiently manage data points unevenly distributed on complex manifolds by mapping them onto a dataset uniformly distributed on a hypercube. This transformation relieves the difficulty of selecting a representative subset on the manifold to a more manageable task of choosing a subset within a hypercube. To ensure the coverage of the design points [21] across lower-dimensional projections, we employ a space-filling design strategy based on maximum projection (MaxPro) [26]. The MaxPro design guarantees that the projection of selected design points onto any subspace maximizes space-filling properties, effectively countering the impact of effect sparsity and thus improving the performance and robustness of the algorithm.

Furthermore, in scenarios involving the fine-tuning of pre-trained models, the unlabeled data pool may include data from classes not recognized by the pre-trained model. In such instances, it is critical to effectively select data from both known and unknown classes to ensure optimal performance. To tackle this challenge, we introduce a re-weighting strategy. This strategy assigns sampling probabilities that reflect not only the distribution of the unlabeled pool but also an updated distribution incorporating insights from the labeled data. By considering both distributions, our approach enables a more informed and effective selection of data from both known and unknown classes, thereby enhancing overall model performance.

We evaluate the SPOT algorithm on three different datasets, specifically targeting the image classification task. The experimental findings demonstrate consistent improvements over baseline methods across these varied datasets and models. In summary, the key contributions of our work are as follows:

- We introduce a novel diversity-based active learning algorithm, named SPOT, which integrates space-filling (SP) designs with the optimal transport (OT) technique. OT efficiently handles

data distributed on complex manifolds, while SP ensures coverage of the design points on lower-dimensional projections.

- We develop a re-weighting strategy designed to enhance the fine-tuning performance by effectively selecting data points from both the known and unknown classes of the pre-trained model.

- We conduct comprehensive experiments across various datasets and models, demonstrating that our SPOT algorithm consistently surpasses the baseline methods. These results provide new perspectives and insights into active learning methods.

2 Related Works

2.1 Active Learning

AL algorithms are generally divided into three main categories: stream-based methods, synthesis-based methods, and pool-based methods. Stream-based AL methods [27, 28, 29, 30] are designed to quickly decide whether to query incoming instances within a data stream. Synthesis-based algorithms [31, 32, 33] generate new instances for querying, rather than selecting from an existing dataset. Pool-based AL methods focus on selecting a specific number of unlabeled instances from an existing pool to optimize learning accuracy. Our study concentrates on pool-based AL methods, which are particularly relevant for DNNs that have access to extensive pools of unlabeled data but limited labeled data. In such scenarios, the importance of each data point for DNNs can be assessed using two main approaches: uncertainty-based and diversity-based methods.

2.2 Uncertainty-based Methods

Uncertainty-based methods [34, 35, 36, 37, 38, 39, 40, 41, 42, 43] aim to query data points with high uncertainty, which suggests that these points are not effectively represented by the pre-trained model. For example, Shannon [35] selects top-k instances with the highest entropy for querying. However, these methods can overlook the structural information of the unlabeled data. As a result, data points belonging to the same category often receive similar uncertainty scores from DNNs, leading to sample bias and the selection of redundant data points [44, 45].

2.3 Diversity-based Methods

This paper primarily focuses on the diversity-based method, which distinguishes itself from uncertainty-based methods by emphasizing the selection of diverse samples that cover the entire sample space, considering distances between all samples. A notable example of this approach is the core-set method, which selects a representative subset by choosing data points that effectively approximate the full dataset's diversity and distribution within a reduced sample space [16, 17, 18]. Despite its strengths, there are situations where the core-set method is outperformed by uncertainty-based methods [46]. One possible explanation is that the core-set method treats each data point equally within the sample space, ignoring the inherent uneven distribution of data across a complex, high-dimensional manifold. Consequently, this method may favor points located in sparse areas, potentially overlooking more critical data points in order to achieve effective coverage. Another limitation of the core-set method arises from the projection of high-dimensional spaces, which can lead to overlapping points in low-dimensional spaces, resulting in information loss and reduced representativeness of the sampled points. These limitations are addressed in our proposed SPOT framework, which enhances the effectiveness of the core-set method.

3 Methodology

We develop a novel AL algorithm named SPOT, which integrates the space-filling design with optimal transport mapping to select a representative subsample. SPOT comprises two main steps. The first step involves linking the feature space to the unit hypercube $[0, 1]^p$, where p is the dimension of data, using the optimal transport technique, enabling the mapping of data points from the complex feature space to a hypercube. In the second step, we employ a space-filling strategy to select the representative subsample that evenly and efficiently covers the hypercube $[0, 1]^p$. The workflow of SPOT is shown in Fig. 2.

3.1 Problem Setup

Using a pre-trained model, an active learning algorithm identifies and selects the most informative data points from a large pool of unlabeled data. These selected data points are then labeled by experts. The newly labeled data is subsequently used to update and refine the model, resulting in enhanced performance.

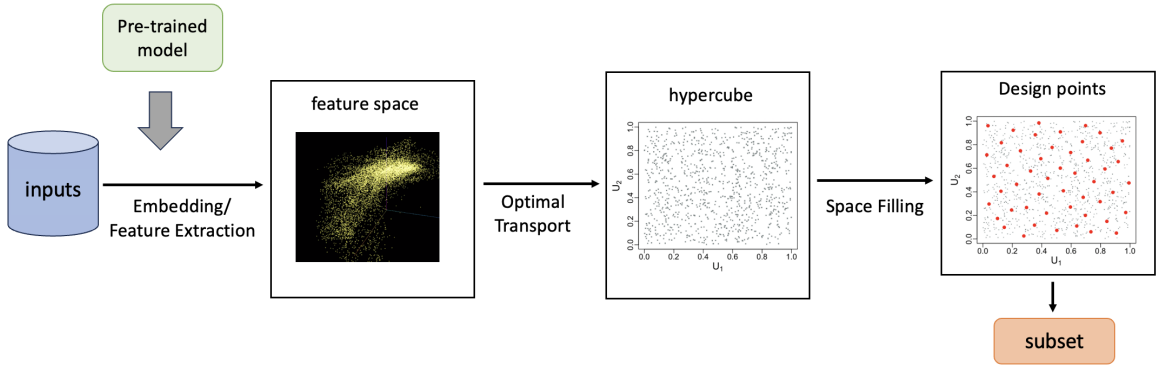

Fig. 2 Workflow of the SPOT algorithm. The inputs are first embedded into the feature space, after which they are mapped into a hypercube via the optimal transport technique. A space-filling strategy is then applied to select the representative subsample (red points) within the hypercube.

Mathematically, consider a pre-trained base model M and an unlabeled pool $\mathcal{D}^U$ containing n unlabeled data points, denoted as $\mathcal{D}^U = \{\mathbf{z}_i \in \mathbb{R}^p\}_{i=1}^n$, where each $\mathbf{z}_i \in \mathbb{R}^p$ represents the p-dimensional covariates of data point i. The objective is to select a fixed-size subset $\mathcal{D}^l = \{\mathbf{z}_j \in \mathbb{R}^p\}_{j=1}^r$ and acquire corresponding labels $\{y_j\}_{j=1}^r$. This subset is chosen to maximize the performance of model M when fine-tuned on $\mathcal{D}^l$. In classification tasks, each y_i is an integer from set $\{1, 2, \ldots, K\}$, representing the class label, where K denotes the number of classes. In regression problems, y_i is a real value.

3.2 SPOT algorithm

3.2.1 Space-filling

To select the subset $\mathcal{D}^l$ that can best represent the whole dataset $\mathcal{D}^U$, we prefer data points that spread evenly in the dataset rather than cluster together. We use star discrepancy, a commonly employed measure, to assess the deviation of a given point set from the uniform distribution. Assuming, without loss of generality, that the unlabeled data points are distributed within the hypercube $[0,1]^p$, our goal is to select a discrete set of data points, $\mathcal{D}^l$, which has the lowest discrepancy.

Given a p-dimensional unit hypercube $[0,1]^p$, let $[0,a) = \prod_{j=1}^p [0,a_j)$ be a hyper-rectangle and $\mathcal{U}_r = \{\mathbf{u}_i\}_{i=1}^r$ be a set of r data points in $[0,1]^p$, the star discrepancy is defined as

$$D^*(\mathcal{U}_r) = \sup_{a\in[0,1]^p} \left| \frac{1}{r}\sum_{i=1}^r 1\{\mathbf{u}_i \in [0,a)\} - \prod_{j=1}^p a_j \right|. \quad (1)$$

The subset $\mathcal{U}_r$ that minimizes D^* is optimal for representing the hypercube space effectively. Several uniform design methods [47] have been proposed to generate such $\mathcal{U}r$. However, these methods are computationally intensive and challenging to apply to datasets with large sample sizes. To reduce the computational load, we employ space-filling design strategies [21, 48], which create $\mathcal{U}r$ with low star discrepancy.

We utilize the Maximum Projection Design (MaxPro), a space-filling strategy, to select a representative subset in $[0,1]^p$. MaxPro helps avoid the suboptimal projections encountered in minimax or maximin distance designs, as illustrated in Fig. 1 (a). In this approach, when data are projected onto a subspace defined by several original dimensions, the distance between points $\mathbf{u}_i$ and $\mathbf{u}_j$ is calculated using the weighted Euclidean distance, defined as:

$$d(\mathbf{u}_i, \mathbf{u}_j; \boldsymbol{\delta}) = \left\{ \sum_{l=1}^p \delta_l (u_{il} - u_{jl})^2 \right\}^{1/2}, \quad (2)$$

where $\mathbf{u}_i = (u_{i1}, \cdots, u_{ip})^T$, $\boldsymbol{\delta} = \{\delta_1, \cdots, \delta_p\}^T$ for $i \in \{1, \cdots r\}$, and $\delta_l = 1$ if dimension l participating in forming the subspace, otherwise $\delta_l = 0$ for $l \in \{1, \cdots, p\}$. We aim to select a subset $\mathcal{U}_r$ that minimizes the projection error across all subspaces, defined as:

$$E\{\phi_k(\mathcal{U}_r; \boldsymbol{\delta})\} = \int_{S_{p-1}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{d^k(\mathbf{u}_i, \mathbf{u}_j; \boldsymbol{\delta})} p(\boldsymbol{\delta}) d\boldsymbol{\delta}, \quad (3)$$

where $S_{p-1} = \left\{\theta : \delta_1, \ldots, \delta_{p-1} \geq 0, \sum_{i=1}^{p-1} \delta_i \leq 1\right\}$ and $k > 0$ is a constant. This ensures optimal representation in each considered subspace. We refer to Joseph et al. [26] for more details.

We propose Algorithm 1 to select the representative subset $\mathcal{U}_r$ within the unit hypercube $[0,1]^p$. This

Algorithm 1 SP Algorithm

Input: The observed sample $\mathcal{D}^U = \{\mathbf{z}_i \in \mathbb{R}^p\}_{i=1}^n$ and budget r.

- Step 1: Scale $\mathcal{D}^U = \{\mathbf{z}_i \in \mathbb{R}^p\}_{i=1}^n$ to $\mathcal{X}^U = \{\mathbf{x}_i \in [0,1]^p\}_{i=1}^n$.
- Step 2: Generate a set of MaxPro space-filling design points $\{\mathbf{u}_j\}_{j=1}^r \in [0,1]^p$.
- Step 3: For $j = 1$ to r,
  Select the nearest neighbor $\mathbf{x}_j$ for $\mathbf{u}_j$ from $\mathcal{X}^U$ using the Euclidean distance.

Output: The selected subset $\{\mathbf{x}_j\}_{j=1}^r$.

approach integrates the space-filling design with a 1-nearest neighbor method similar to Zhang et al. [49]. First, we scale the original sample $\mathcal{D}^U$ to $\mathcal{X}^U$, ensuring it is distributed within $[0,1]^p$. We then generate MaxPro design points within this space. For each design point, denoted as $\mathbf{u}_j \in [0,1]^p$, we identify its nearest neighbor $\mathbf{x}_j \in \mathcal{X}^U$. This neighboring point $\mathbf{x}_j$ is the data point we select to fine-tune the model.

3.2.2 Optimal Transport

For any $\mathcal{D}^U = \{\mathbf{z}_j \in \mathbb{R}^p\}_{j=1}^r$, Algorithm 1 can be applied following a simple scaling step. Nonetheless, challenges arise when the data points are non-uniformly distributed across the sample space. Employing the MaxPro space-filling design method under these conditions often leads to suboptimal outcomes. Firstly, as illustrated in Fig. 3 (a), Algorithm 1 tends to select the subset that overly represents data points from sparse areas. Secondly, for data points that are non-uniformly distributed in the sample space, utilizing a uniformly distributed space-filling design set to locate the nearest neighbor may not be reasonable. This is because even its nearest neighbor can still be significantly distant, making this approach ineffective.

We apply the optimal transport (OT) technique [22, 50] to transfer the dataset $\mathcal{D}^U$, which is unevenly distributed on a complex manifold, into a uniformly distributed dataset within a unit hypercube $[0,1]^p$. The transformation simplifies the challenging task of selecting a representative subset from the manifold to selecting one from a dataset uniformly distributed in a hypercube. Consequently, the effectiveness of Algorithm 1 is fully demonstrated, as shown in Fig. 3 (b). We observe that the selected data points are more concentrated to the true distribution, and it is robust to

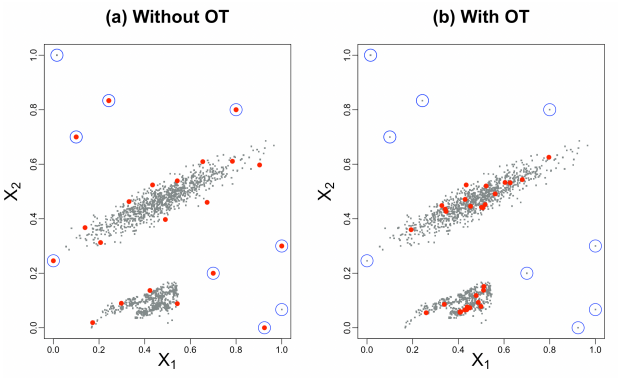

Fig. 3 Power of OT on the unevenly distributed sample (grey points). The points from the sparse areas are considered as outliers (circled in blue). a) Subset (red points) selected by applying algorithm 1 directly. b) Subset (red points) selected by applying algorithm 1 after the optimal transport method.

this non-uniformly distribution with outliers.

Assume μ is the probability measure on space $X \in \mathbb{R}^p$, the domain of the random variable, and ν is the uniform probability measure on $Y = [0,1]^p$. Let $T : X \to Y$ be a transport map that transports $\mu \in \mathcal{P}(X)$ to $\nu \in \mathcal{P}(Y)$, where $\mathcal{P}(\cdot)$ is the set of probability measures on $(\cdot)$. T is defined such that

$$\nu(B) = \mu\left(T^{-1}(B)\right), \tag{4}$$

for all ν-measurable sets B. As shorthand we write $\nu = T_{\#}\mu$ if Eq. (4) is satisfied. The focus is primarily on the cost of transporting μ to ν. Specifically, let $c : X \times Y \to [0, +\infty]$ be a cost function, where $c(\mathbf{x}, \mathbf{y})$ measures the cost of transporting one unit of mass from $\mathbf{x} \in X$ to $\mathbf{y} \in Y$. The objective is to search the optimal transport map T^* that minimizes

$$\mathbb{M}(T) = \int_X c(\mathbf{x}, T(\mathbf{x}))d\mu(\mathbf{x}), \tag{5}$$

over μ-measurable maps $T : X \to Y$ subject to $\nu = T_{\#}\mu$.

To obtain the desired optimal transport map that maps the observed sample to be uniformly distributed on $[0,1]^p$, a synthetic sample, $\mathcal{U}_n = \{\mathbf{u}_i\}_{i=1}^n$, uniformly distributed on $[0,1]^p$ is first generated. Subsequently, T^*, mapping from the observed sample to $\mathcal{U}_n$ is calculated. This mapping can be approximated using projection-based methods [25, 51], which simplify the estimation of a p-dimensional optimal transport map by addressing it through a sequence of one-dimensional subproblems. These subproblems, involving the calculation of one-dimensional optimal transport maps between projected samples, are readily solved using sorting algorithms. The set $\mathcal{U}_r$ is then selected according to Eq. 1 based on space-filling designs.

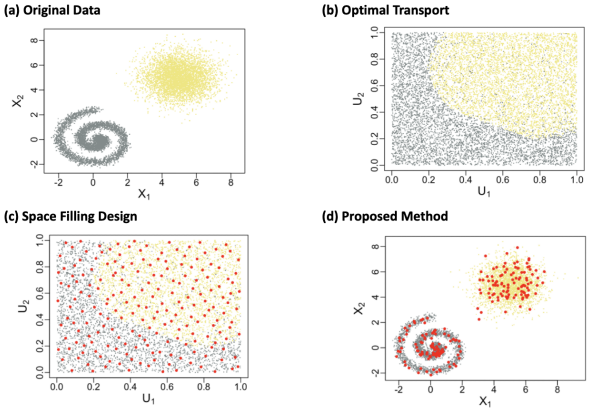

Fig. 4 Toy example of the proposed SPOT algorithm. (a) original data consisting of two classes, distinguished by different colors; (b) optimal transport maps the original data to the synthetic data uniformly distributed on the $2d$ unit hypercube $[0,1]^2$; (c) generated space-filling design points (red points) covering the unit hypercube $[0,1]^2$; (d) subset of the original data (red points) mapped to the selected synthetic data.

Algorithm 2 Naive-SPOT Algorithm

Input: $\mathcal{D}^U$, $\mathcal{D}^L$, and budget r.

- Step 1: Generate a synthetic sample $\mathcal{U}_n = \{\mathbf{u}_i\}_{i=1}^n$ uniformly distributed on the unit hypercube $[0,1]^p$.
- Step 2: Calculate the optimal transport map T^* that maps $\mathcal{D}^U = \{\mathbf{z}_i \in \mathbb{R}^p\}_{i=1}^n$ to $\mathcal{U}_n$.
- Step 3: Generate the MaxPro space-filling design points $\{\mathbf{u}_j\}_{j=1}^r$ = SP($\mathcal{U}_n$, r).
- Step 4: Achieve the subset $\mathcal{D}^l = \{\mathbf{z}_j\}_{j=1}^r$ mapped to $\{\mathbf{u}_j\}_{k=1}^r$ by T^*.

Output: Selected subset $\mathcal{D}^l$.

The observed samples transported to $\mathcal{U}_r$ by T^* form the targeted subsample. This procedure is outlined in Algorithm 2. The selected subset is subsequently annotated with expert knowledge and utilized to refine the current model, M.

We further illustrate algorithm 2 using a toy example as shown in Fig. 4. We generate two distinct classes of random samples, each consisting of 3,000 points. The first class is sampled from a normal distribution $N\left(\begin{pmatrix} rsin(2r) \\ rcos(2r) \end{pmatrix}, \begin{pmatrix} 0.4^2 & 0 \\ 0 & 0.4^2 \end{pmatrix}\right)$, where $r \sim Unif(0, 2\pi)$, and the second from $N\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$, as displayed in Fig. 4 (a). Following this, we map the generated data to a synthetic dataset uniformly distributed on $[0,1]^2$ as per step 2 in Algorithm 2. Figure 4 (b) confirms that data from the same class remain spatially close even after the OT step, validating the logic of the subsequent selection procedure (SP). The SP design points are marked in red in Fig. 4 (c). The subsample corresponding to these design points, marked in red in Fig. 4 (d), form the desired subsample.

3.2.3 Down-weight

Algorithm 2 is designed to select a small subset of data that effectively represents the entire unlabeled dataset. This is particularly useful when the unlabeled data comes from classes that differ from those in the base dataset $\mathcal{D}_0^L$, which is used to train the base model. However, in practice, $\mathcal{D}^U$ may also contain data points belonging to the same classes as $\mathcal{D}_0^L$, which the base model already distinguishes well. In these cases, we prefer to reduce the probability of selecting such data points. To address this issue, we introduce a down-weighting method that adjusts the input to our selection procedure.

To illustrate our method, consider a scenario involving two classes. Assume the unlabeled dataset $\mathcal{D}^U = \{\mathbf{z}_i \in \mathbb{R}^p\}_{i=1}^n$ contains two classes: $\mathcal{C}_1^U = \{\mathbf{z}_j : j \in I_1\}$ and $\mathcal{C}_2^U = \{\mathbf{z}_k : k \in I_2\}$, where $I_1 \cup I_2 = \{1, \cdots, n\}$, $I_1 \cap I_2 = \varnothing$, $|I_1| = m_1$, and $|I_2| = m_2$. Furthermore, suppose that $\mathcal{C}_1^U$ contains the same classes as the base labeled dataset $\mathcal{D}_0^L = \{\mathbf{x}_i\}_{i=1}^N$, where N is the sample size of $\mathcal{D}_0^L$. We denote the sampling probabilities for data points in $\mathcal{C}_1^U$, $\mathcal{C}_2^U$, and $\mathcal{D}_0^L$ by p_{1j} (for $j \in I_1$), p_{2k} (for $k \in I_2$), and p_{0i} (for $i = 1, \cdots N$), respectively.

According to the principle of OT and SP, the proportion of the selected subset from each class should be proportional to the sample size of each class. Therefore, when incorporating the base dataset $\mathcal{D}_0^L$, the probability of selecting each data point into the subset is given by:

$$\hat{p}_{0i} = \frac{n}{n + m_1 + m_2} \times \frac{p_{0i}}{\sum_{i=1}^n p_{0i} + \sum_{i=1}^{m_1} p_{1i}}, \quad (6)$$

$$\hat{p}_{1j} = \frac{m_1}{n + m_1 + m_2} \times \frac{p_{1j}}{\sum_{i=1}^n p_{0i} + \sum_{i=1}^{m_1} p_{1i}}, \quad (7)$$

$$\hat{p}_{2k} = \frac{m_2}{n + m_1 + m_2} \times \frac{p_{2k}}{\sum_{i=1}^{m_2} p_{2i}}, \quad (8)$$

where $\hat{p}_{0i}$, $\hat{p}_{1j}$, $\hat{p}_{2k}$ represent the adjusted sampling probability for each data point in $\mathcal{D}_0^L$, $\mathcal{C}_1^U$, $\mathcal{C}_2^U$, respectively. Without including the base dataset, the

adjusted probabilities are as follows:

$$\tilde{p}_{1j} = \frac{m_1}{m_1 + m_2} \times \frac{p_{1j}}{\sum_{i=1}^{m_1} p_{1i}}, \tag{9}$$

$$\tilde{p}_{2k} = \frac{m_2}{m_1 + m_2} \times \frac{p_{2k}}{\sum_{i=1}^{m_2} p_{2i}}. \tag{10}$$

For any $\mathbf{z}_j \in \mathcal{C}_1^U$ and $\mathbf{z}_k \in \mathcal{C}_2^U$, without the base dataset, the ratio of the selection probability is given by

$$k_0 = \frac{\hat{p}_{1j}}{\hat{p}_{2k}} = \frac{p_{1j} \sum_{i=1}^{m_2} p_{2i}}{p_{2k} \sum_{i=1}^{m_1} p_{1i}}. \tag{11}$$

However, when including the base dataset, this ratio becomes

$$k_1 = \frac{\tilde{p}_{1j}}{\tilde{p}_{2k}} = \frac{p_{1j} \sum_{i=1}^{m_2} p_{2i}}{p_{2k} \left(\sum_{i=1}^{n} p_{0i} + \sum_{i=1}^{m_1} p_{1i}\right)}, \tag{12}$$

which is lower than k_0. Thus, we effectively decrease the probability of selecting data points from classes that are already well represented. Further details of this method under general conditions are outlined in the Algorithm 3.

Algorithm 3 SPOT Algorithm

Input: $\mathcal{D}^U$, $\mathcal{D}_0^L$, and budget r.

- Step 1: Randomly select a subset $\mathcal{D}_{\text{sub}}^L$ from $\mathcal{D}_0^L$ with the size of the sample order with $\mathcal{D}^U$.
- Step 2: Form $\mathcal{D}_{\text{new}} = \mathcal{D}^U \cup \mathcal{D}_{\text{sub}}^L$.
- Step 3: Pass $\mathcal{D}_{\text{new}}$ into the Naive-SPOT algorithm to select a subset for labeling:

$$\mathcal{D}^l = \text{Naive-SPOT}(\mathcal{D}_{\text{new}}, r)$$

.

  - If the selected subset contains samples from $\mathcal{D}_0^L$, no budget is used for these alreadylabeled data points.
  - The saved budget can either be preserved or used to annotate additional samples.

Output: Selected subset $\mathcal{D}^l$.

In the first step of the SPOT Algorithm 3, simple random sampling is employed to select a subset from the labeled base dataset, $\mathcal{D}_0^L$. To enhance the performance of the SPOT algorithm, the potential for incorporating more advanced sampling techniques [52, 53, 54] can be further investigated.

4 Experiments

This section provides an overview of the datasets and algorithms to be employed in our experiments, followed by an experimental analysis. We perform a thorough evaluation of SPOT across multiple classification tasks utilizing various models. Furthermore, we conduct a sensitivity analysis to assess the effects of several critical parameters on the performance of the SPOT algorithm.

Table 1 Datasets used in the experiments.

Dataset	m_1	m_2	n_1	n_2	Model
CIFAR-10	7	6	30,500	19,500	ViT
Agri-ImageNet	3	8	3,149	1,491	ViT
MNIST	7	6	36,781	23,219	CNN

m_1: # classes in $\mathcal{D}_0^L$; m_2: # classes in $\mathcal{D}^U$;
n_1: # images in $\mathcal{D}_0^L$; n_2: # images in $\mathcal{D}^U$.

4.1 Baselines

To validate the performance of our approach, we compare it against a number of baselines:

- **Coreset:** Following the K-center algorithm (K is equal to the budget) developed in [16] to select $\mathcal{D}^l$. We use Gurobi [55] to iteratively solve the integer program.
- **BADGE:** Batch Active learning by Diverse Gradient Embeddings according to [56].
- **K-means:** Partitioning $\mathcal{D}^U$ into K clusters (K is equal to the budget) according to [57] and take the cluster centroids as $\mathcal{D}^l$.
- **Random:** Selecting the subset $\mathcal{D}^l$ uniformly at random from $\mathcal{D}^U$.
- **Least Confidence (LC) :** Selecting $\mathcal{D}^l$ for which the pre-trained model M is least confident in class assignment.
- **ALBL:** Active Learning by Learning. A bandit-style meta-active learning algorithm that selects between Coreset and LC at every round [58].
- **GEFD:** low generalized empirical F-discrepancy (GEFD) data-driven subsampling method according to [59].

4.2 Size of the Budget

Different from many previous AL studies that allocate a large budget for their experiments, we focus on the scenarios where budget B is very limited. This focus mirrors situations where labeling is extremely expensive, as is often the case in fields such as medical imaging. Specifically, for the situation that $\mathcal{D}^U$ contains tens of thousands of data points, we limit the size of the budget to the order of tens, i.e., the few-shot scenario [60, 61].

4.3 Dataset

Agri-ImageNet: The Agri-ImageNet dataset [62] contains two parent classes including fruits (with 11

sub-classes) and vegetables (with 4 sub-classes).

MNIST: MNIST [63] is a dataset of handwritten digit images with a training set of $60,000$ examples and a test set of $10,000$ examples. Each example is a 28×28 grayscale image, associated with a label of 10 classes.

CIFAR-10: The CIFAR-10 dataset [64] consists of a training set of $50,000$ examples and a test set of $10,000$ examples. Each example in the dataset is a 32×32 color image, spanning 10 different classes of objects such as animals and vehicles. These classes include airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks, each equally represented in the dataset.

4.4 Implementation Details

We briefly introduce some important implementation details of our experiments. We refer to the Experiments section in the Supplementary Materials for more detailed settings about the models.

Dataset settings: For all datasets, we randomly separate them into the base dataset and the novel dataset. Model M is pre-trained on the base set, and the active learning algorithms are applied to the novel set. The base dataset is randomly divided into training (80%) and test (20%) splits. For the novel dataset, all samples except those actively selected for fine-tuning are used as the test split. Image pre-processing steps are also applied. Specifically, for the training dataset, Rand-Augment [65], Random Erasing [66], and RandomResizeCrop is applied for data augmentation. For the test dataset, images are only resized and center-cropped.

In table 1, we list some basic information about the three datasets.

Model settings: We consider two different model structures. For Agri-ImageNet and CIFAR-10 dataset, we apply the Vision Transformer (ViT) [2] model in the experiments. The ImageNet-1k pre-trained model is firstly trained on the base dataset with the vanilla ViT. We adopt an AdamW optimizer with 300 epochs using a cosine decay learning rate scheduler and 5 epochs of linear warm-up. For the MNIST dataset, a Convolutional Neural Network with two sequential layers and one fully connected layer is applied.

Feature extraction: For the distance-based methods (Coreset, K-means), we follow the instructions in [16] to define the distance metric. Specifically, we use the l_2 distance between the final fully connected layers as the distance. For SPOT, since the properties of space-filling designs are restricted to a relatively low dimension, we further apply a simple Autoencoder and PCA step to reduce the dimension. This feature extraction procedure is solely used for selecting $\mathcal{D}^l$ and will not be applied in the subsequent model fine-tuning step.

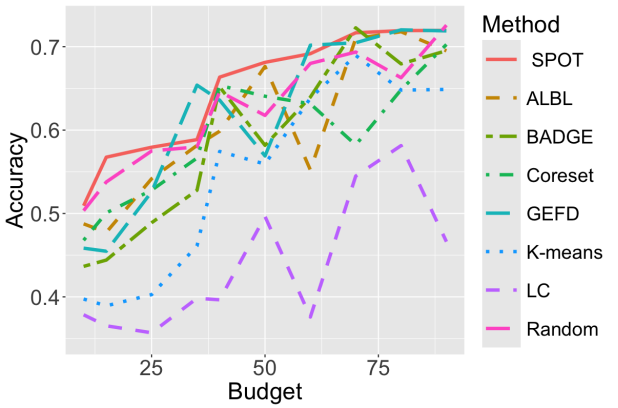

Fig. 5 The image classification accuracy given the budgets (number of training samples) on the CIFAR-10 dataset with the Vision Transformer model. Accuracies are plotted versus different budget r (x axis).

For all the active learning algorithms with randomness, we run them with three random seeds and use the median accuracy as a metric.

4.5 Results

Figure 5, Fig. 6, and Fig. 7 show the results of classification accuracies versus different budget r. Three significant observations can be made from these results. First, it is observed that as the budget increases, the accuracy of all methods generally exhibits an upward trend. Although there may be slight drops in accuracy at certain points, such as when $r = 50$ for the SPOT algorithm, the overall trend remains positive. Notably, the proposed SPOT algorithm consistently outperforms the other methods for both datasets in most cases with a few exceptions that GEFD achieves marginally better accuracies. These findings align with the statements and demonstrations provided in the methodology section and the accompanying toy examples. They reinforce the notion that the subset selected by the SPOT algorithm better represents the observed sample space compared to the subsets selected by the other four methods.

Second, we note that even with a significantly limited budget (specifically, a budget controlled to be under 100), DNNs can still achieve good performance by taking advantage of active learning algorithms. Utilizing the SPOT algorithm, the classification accuracy on both datasets reaches 70%. This highlights

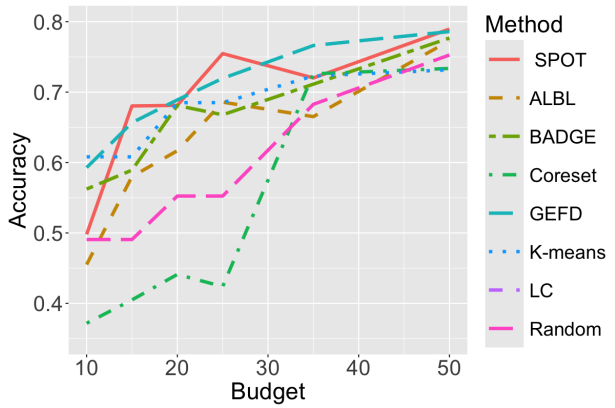

Fig. 6 The image classification accuracy given the budgets (number of training samples) on the Agri-ImageNet dataset with the Vision Transformer model. Accuracies are plotted versus different budget r (x axis).

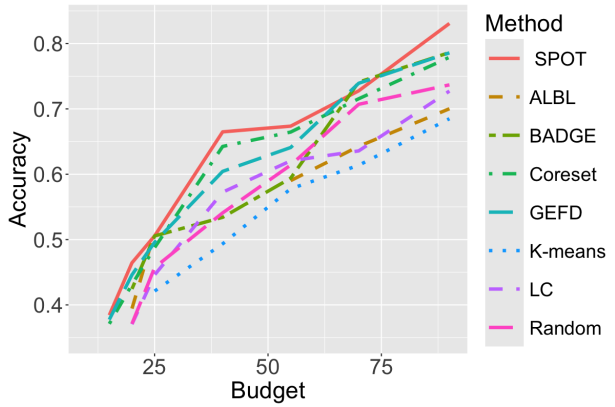

Fig. 7 The image classification accuracy given the budgets (number of training samples) on the MNIST dataset with the CNN model. Accuracies are plotted versus different budget r (x axis).

the efficacy of SPOT in maximizing performance even under resource constraints.

Third, we observe that the accuracy of all methods gradually approaches a fixed value, differing only in their convergence rates. This behavior is expected since, as the training sample size increases, the distinctions between various active learning methods diminish until they become negligible. For instance, the convergence rate of the star discrepancy for space-filling design points is of the order $O\left(\log(r)^p/r\right)$, while the convergence rate for uniformly random sampling is of the order $\mathcal{O}(log(log(r))/\sqrt{r})$ [67], which is significantly slower than $O\left(\log(r)^p/r\right)$. However, as the budget r goes to infinity, even uniformly random sampling will perform well. One exception is the performance of the entropy-based method LC in the Agri-ImageNet dataset. As r increases, its accuracy barely changes. This may come from the fact that DNNs tend to give similar uncertainties to the data points belonging to the same class.

4.6 Computational Time

Although the expensive labeling procedure constitutes a significant cost in active learning algorithms, it is also essential to consider the computational time required by the proposed SPOT algorithm. Typically, the pre-trained model used in active learning is not counted as part of the computational cost since it is trained on large benchmark datasets like ImageNet. Thus, the computational time for the model-building procedure consists of two main parts: 1) selecting the subset $\mathcal{D}^l$ for annotation, and 2) the fine-tuning process to adapt the pre-trained model to the novel dataset. The subset selection step is performed on a Mac with a 10-Core M1 Max processor and 32 GB memory, utilizing the CPU. On the other hand, the fine-tuning process is executed on an NVIDIA Tesla V100 Tensor Core. We list the computational time of step 1 in seconds in Table 2 and the computational time for step 2 in hours in Table 3.

Table 2 shows that the computational time required by different subset selection methods varies greatly, but overall this step can usually be completed within several minutes. A running time of zero here indicates that the execution is completed in less than one second. While the fine-tuning step is more time-consuming, requiring several hours to fine-tune the ViT model.

Table 2 Median Computational time (sec)

	Subset selection part							
Method	SPOT	Coreset	K-means	Random	LC	ALBL	BADGE	GEFD
Agri-ImageNet	23	50	416	0	119	201	223	1
MNIST	174	6	1300	0	39	205	556	0
CIFAR-10	132	83	1081	0	54	371	368	1

Table 3 Median Computational time (hour)

	Fine-tuning part							
Method	SPOT	Coreset	K-means	Random	LC	ALBL	BADGE	GEFD
Agri-ImageNet	11.57	12.25	14.55	11.38	14.37	19.07	19.06	16.52
MNIST	0.17	0.18	0.15	0.18	0.08	0.07	0.13	0.12
CIFAR-10	3.08	3.89	3.43	3.05	2.55	1.71	3.51	3.20

4.7 Parameter Sensitivity

To assess the impact of parameterization changes in the dimension reduction phase on classification performance, we conducted experiments using the

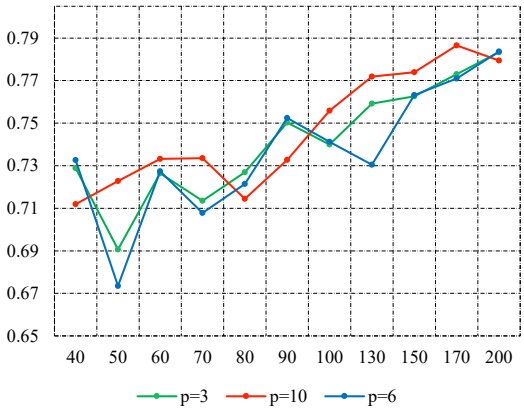

Fig. 8 Results of the parameter sensitivity test for the dimension reduction part.

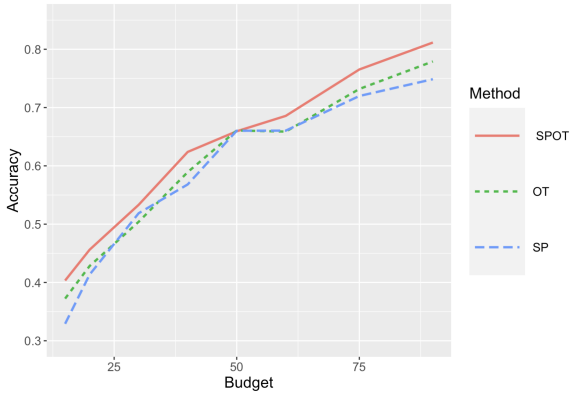

Fig. 9 The image classification accuracy given the budgets (number of training samples) on the MNIST dataset. Accuracies are plotted versus different budget r (x axis).

CIFAR-10 dataset. Specifically, we analyzed the effects of modifying the number of the first principal component, denoted as p. The results, as depicted in Fig. 8, consistently demonstrate stable classification performance across various values of p. This finding suggests that the performance remains robust and unaffected by changes in the specific values of p. More details of experiments can be found in Appendix A.4.

4.8 Ablation Study

We conduct experiments to assess the influence of space-filling designs and OT individually. Using the MNIST dataset as an example, we compare SPOT with the following methods: (1) OT, which applies optimal transport with a simple random Latin hypercube design [68] instead of the proposed MaxPro space-filling design; and (2) SP, which uses the MaxPro space-filling design without OT. Figure 9 shows the classification accuracies of these three methods at varying budget levels r. The results demonstrate that SPOT consistently achieves higher classification accuracy compared to both OT and SP, with the performance gap increasing at higher budget levels.

5 Conclusion and Discussion

In this paper, we introduce a novel active learning framework that combines space-filling (SP) designs and optimal transport (OT) to effectively select representative subsets that capture the underlying distribution of the entire dataset. In particular, our design remedies the limitations in core-set-based methods from the uneven distribution density of data points and ineffective projection onto sub-spaces. Through extensive experiments on three diverse datasets using various models, we demonstrate the superiority of our proposed methods compared to the baseline approaches. The results highlight the effectiveness and robustness of our framework. As part of our future work, we aim to apply the SPOT framework to other scenarios, including medical imaging applications and other data modalities such as text, time series, and videos. This will allow us to explore the potential benefits and practicality of our approach in broader domains.

The computational cost of SPOT depends on both the OT step and the SP step. Traditional linear programming algorithms for solving OT problems have a computational complexity of $O\left(n^3 \log(n)\right)$. Additionally, the MaxPro design step has a complexity of $O\left(n^2 \cdot p\right)$, where n is the sample size and p is the data dimension. For large-scale datasets, such as medical imaging data, the high computational cost of OT poses a significant challenge to implementing SPOT. Fortunately, efficient OT algorithms, such as the Sinkhorn algorithm, have been developed to significantly reduce computational time. Empirical studies demonstrate that the Sinkhorn algorithm, with a complexity of $O\left(n^2 \log(n)\right)$, can solve OT problems reliably and efficiently for datasets with $n \approx 10^4$ [69]. Furthermore, under sparsity assumptions, the computational cost can be further reduced, with efficiency demonstrated on datasets as large as $n \approx 10^6$ [70]. Thus, the SPOT algorithm remains feasible and practical for most applications, even with large-scale datasets.

Acknowledgment

This work was partially supported by the U.S. National Science Foundation [DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809] and the National Institutes of Health [NIH R01GM152814].

References

[1] Huiyan Jiang, Zhaoshuo Diao, Tianyu Shi, Yang Zhou, Feiyu Wang, Wenrui Hu, Xiaolin Zhu, Shijie Luo, Guoyu Tong, and Yu-Dong Yao. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Computers in Biology and Medicine*, 157:106726, 2023.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Luyang Fang, Yongkai Chen, Wenxuan Zhong, and Ping Ma. Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification. In *Forty-first International Conference on Machine Learning*, 2024.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[7] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10, 2020.

[8] Brian R Bartoldson, Bhavya Kailkhura, and Davis Blalock. Compute-efficient deep learning: Algorithmic trends and opportunities. *Journal of Machine Learning Research*, 24(122):1–77, 2023.

[9] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V Vasilakos. Privacy and security issues in deep learning: A survey. *IEEE Access*, 9:4566–4593, 2020.

[10] Burr Settles. Active learning literature survey. 2009.

[11] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9): 1–40, 2021.

[12] Chunhui Zhao, Boao Qin, Shou Feng, Wenxiang Zhu, Weiwei Sun, Wei Li, and Xiuping Jia. Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning. *IEEE Transactions on Image Processing*, 32:3606–3621, 2023.

[13] Saed Rezayi, Haixing Dai, Zhengliang Liu, Zihao Wu, Akarsh Hebbar, Andrew H Burns, Lin Zhao, Dajiang Zhu, Quanzheng Li, Wei Liu, et al. Clinicalradiobert: Knowledge-infused few shot learning for clinical notes named entity recognition. In *International Workshop on Machine Learning in Medical Imaging*, pages 269–278. Springer, 2022.

[14] Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

[15] Jeff M Phillips. Coresets and sketches. In *Handbook of discrete and computational geometry*, pages 1269–1288. Chapman and Hall/CRC, 2017.

[16] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[17] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. *Advances in neural information processing systems*, 32, 2019.

[18] 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.

[19] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies*. Springer Science & Business Media, 2009.

[20] Mark E Johnson, Leslie M Moore, and Donald Ylvisaker. Minimax and maximin distance designs. *Journal of statistical planning and inference*, 26(2):131–148, 1990.

[21] CF Jeff Wu and Michael S Hamada. *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons, 2011.

[22] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[23] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[24] Jingyi Zhang, Ping Ma, Wenxuan Zhong, and Cheng Meng. Projection-based techniques for high-dimensional optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1587, 2022.

[25] Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] V Roshan Joseph, Evren Gul, and Shan Ba. Maximum projection designs for computer experiments. *Biometrika*, 102(2):371–380, 2015.

[27] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

[28] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[29] Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 20, 2007.

[30] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285:181–203, 2014.

[31] Dana Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.

[32] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[33] Miriam Huijser and Jan C van Gemert. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE international conference on computer vision*, pages 5286–5295, 2017.

[34] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[35] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[36] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021.

[37] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9433–9443, 2020.

[38] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

[39] Colin Campbell, Nello Cristianini, Alex Smola, et al. Query learning with large margin classifiers. In *ICML*, volume 20, page 0, 2000.

[40] Ashish Kapoor, Gang Hua, Amir Akbarzadeh, and Simon Baker. Which faces to tag: Adding prior constraints into active learning. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1058–1065. IEEE, 2009.

[41] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

[42] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[43] Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel. Active learning with support vector machines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):313–326, 2014.

[44] Michael Bloodgood and K Vijay-Shanker. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. *arXiv preprint arXiv:1409.4835*, 2014.

[45] Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

[46] 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.

[47] Kaitai Fang, Min-Qian Liu, Hong Qin, and Yong-Dao Zhou. *Theory and application of uniform experimental designs*, volume 221. Springer, 2018.

[48] V Roshan Joseph. Space-filling designs for computer experiments: A review. *Quality Engineering*, 28(1):28–35, 2016.

[49] Jingyi Zhang, Cheng Meng, Jun Yu, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. An optimal transport approach for selecting a representative subsample with application in efficient kernel density estimation. *Journal of Computational and Graphical Statistics*, (just-accepted):1–26, 2022.

[50] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86), 2017.

[51] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[52] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

[53] Ping Ma, Xinlian Zhang, Xin Xing, Jingyi Ma, and Michael Mahoney. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1026–1035. PMLR, 2020.

[54] Jia Guo, Hongxiang Gu, and Miodrag Potkonjak. Efficient image sensor subsampling for dnn-based image classification. In *Proceedings of the International Symposium on Low Power Electronics and Design*, pages 1–6, 2018.

[55] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2018.

[56] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[57] David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178, 2010.

[58] Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[59] Mei Zhang, Yongdao Zhou, Zheng Zhou, and Aijun Zhang. Model-free subsampling method based on uniform designs. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[60] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[61] Michael Fink. Object classification from a single example utilizing class relevance metrics. *Advances in neural information processing systems*, 17, 2004.

[62] Yuzhong Chen, Zhenxiang Xiao, Lin Zhao, Lu Zhang, Haixing Dai, David Weizhong Liu, Zihao Wu, Changhe Li, Tuo Zhang, Changying Li, et al. Mask-guided vision transformer (mg-vit) for few-shot learning. *arXiv preprint arXiv:2205.09995*, 2022.

[63] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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

[65] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[66] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[67] Kai-Lai Chung. An estimate concerning the kolmogroff limit distribution. *Transactions of the American Mathematical Society*, 67(1):36–50, 1949.

[68] Michael Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

[69] Jingyi Zhang, Ping Ma, Wenxuan Zhong, and Cheng Meng. Projection-based techniques for high-dimensional optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1587, 2023.

[70] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. *Advances in neural information processing systems*, 32, 2019.

Luyang Fang received her M.S. degree in statistics from the University of Wisconsin-Madison in 2021. She is currently pursuing a Ph.D. in statistics at the University of Georgia. Her research interests include deep learning, non-parametric methods, and big data analytics.

Cheng Meng received the PhD degree from the Department of Statistics, University of Georgia, in 2020. He is an assistant professor (tenure-track) with the Institute of Statistics and Big Data, Renmin University of China. His research interests include numerical linear algebra, optimal transport problems, sufficient dimension reduction, nonparametric statistics, and machine learning.

Lin Zhao received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Computing, University of Georgia, Athens, GA, USA, under the supervision of Prof. Tianming Liu. His current research interests include deep learning and medical image analysis.

Tao Wang received the M.S. degree in Operations Research from Georgia Institute of Technology in 2021. He is currently pursuing a Ph.D. in statistics at the University of Georgia. His current research interests include deep learning, functional regression and large language model.

Tianming Liu is a Distinguished Research Professor and a Full Professor of Computer Science at The University of Georgia. Dr. Liu's research interests are brain imaging, computational neuroscience, brain-inspired artificial intelligence, and artificial general intelligence. Dr. Liu has published 400+ research papers on these topics, his Google citation is over 16,000+, and his H-index is 66. Dr. Liu is the recipient of NIH Career Award and NSF CAREER Award. Dr. Liu serves on the editorial boards of multiple international journals including IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Medical Imaging, Medical Image Analysis, IEEE Transactions on Cognitive and Developmental Systems, IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE Reviews in Biomedical Engineering, and IEEE Journal of Biomedical and Health Informatics. Dr. Liu is a Fellow of AIMBE (American Institute of Medical and Biological Engineering) and was the General Chair of MICCAI 2019.

Wenxuan Zhong earned her B.S. in Statistics from Nankai University and a Ph.D. from Purdue University. After a postdoctoral fellowship in statistics and systems biology at Harvard, she served as an assistant professor at UIUC from 2007-2013. Since 2013, she has been with the Department of Statistics at UGA, where she is now the Georgia Athletic Association Professor. Dr. Zhong's research focuses on the statistical methodology and theory development to face the striking new phenomena emerged under the big data regime. Over the past few years, Dr. Zhong has established diverse extramurally funded research programs to overcome the computational and theoretical challenges arise from the big data analysis. The basic statistical researches are successfully applied in modern genomic, epigenetic, metagenomics, text-mining and chemical sensing researches.

Ping Ma graduated with a B.S. in statistics from Nankai University and a Ph.D in statistics from Purdue University. He then conducted postdoctoral research at Harvard University. Since 2005, he has served as an Assistant Professor and then Associate Professor at the University of Illinois at Urbana-Champaign. He is currently a Distinguished Research Professor at the University of Georgia. Dr. Ma's main research areas include the statistical theory and methods for very large samples, the development and application of data-driven nonparametric inverse modeling methods, and the development of gene regulatory network analysis.

Supplementary Material for "SPOT: An Active Learning Framework for Deep Neural Networks"

A Experiments

In this section, we provide detailed information of our experiments.

A.1 Dataset Settings:

A.1.1 Splitting the dataset:

We partitioned each dataset into two subsets: the base dataset and the novel dataset. The pre-trained model was trained using the base dataset, while the active learning algorithm was applied to the novel dataset. In the case of the CIFAR-10 and MNIST datasets, the novel dataset comprises both classes that are already present in the base dataset and additional classes that are not included in the base dataset. For the Agri-ImageNet dataset, all the classes in the novel dataset are entirely new."

CIFAR-10: We design all data samples belonging to four classes (airplane, automobile, bird, cat) and randomly allocate 70% of the data from three classes (deer, dog, frog) to form the base dataset. Subsequently, we assign the remaining 30% of the three classes (deer, dog, frog) along with all data samples from three classes (horse, ship, truck) as the novel dataset.

Agri-ImageNet: The base dataset contains three classes (Chinee apple, maize, and tomato), while the novel dataset contains 12 classes (apple, fuji apple, golden delicious apple, melrose apple, apple tree, avocado, capsicum, lettuce, mango, orange, rockmelon, and strawberry).

MNIST: Similar to the CIFAR-10 dataset, we set all data from four classes (digit 0-3) and randomly select 70% of the data from three classes (digit 4-6) as the base dataset. We then set the rest of the data, i.e. 30% of three classes (digit 4-6) and all data from three classes (digit 7-9), as the novel dataset.

A.1.2 Dataset settings:

For both datasets, the base dataset is randomly split into training/testing with 80%/20%. The remaining data in the novel dataset except for the actively selected few-shot samples is the test split of the novel dataset. Image pre-processing steps are also applied. Specifically, for the training dataset, Rand-Augment [65], Random Erasing [66], and RandomResizeCrop to 32×32 for CIFAR-10, to 224×224 for Agri-ImageNet are applied for data augmentation. For the test dataset, images are only resized and center cropped to 32×32 for CIFAR-10, and 224×224 for Agri-ImageNet.

A.2 Model Settings

ViT model for CIFAR10: We use the Vision Transformer (ViT) [2] model in the experiments. The ImageNet-1k pre-trained model is firstly trained on the base dataset with the vanilla ViT. We adopt an AdamW optimizer with 300 epochs using a cosine decay learning rate scheduler and 5 epochs of linear warm-up. Then, we fine-tune the model on the few-shot samples in the novel dataset. We keep the same settings of regular training except for the epochs to 200.

ViT model for Agri-ImageNet: We use the Vision Transformer (ViT) [2] model in the experiments. The ImageNet-1k pre-trained model is firstly trained on the base dataset with the vanilla ViT. We adopt an AdamW optimizer with 100 epochs using a cosine decay learning rate scheduler and 5 epochs of linear warm-up. Then, we fine-tune the model on the few-shot samples in the novel dataset. We keep the same settings as regular training.

CNN model for MNIST: We use a Convolutional Neural Network with two sequential layers and three fully connected layers. The CNN model is first trained on the base dataset. We adopt an Adam optimizer with 100 epochs and 5 epochs of linear warm-up. Then, we fine-tune the model on the few-shot samples in the novel dataset. We keep the same settings of regular training except for the epochs to 300.

A.3 Feature extraction

For particularly high-dimensional data such as images, it is not reliable or even feasible for us to use the original high-dimensional data for analysis. Thus, a feature extraction step, which has the ability to extract low-dimensional features that can preserve the most relevant information from the original dataset and discard the redundant information, is desired before applying the active learning algorithms. For the classification problems, since the pre-trained model itself has the ability to extract important features required to distinguish classes, we take advantage of it to finish the feature extraction step.

ViT model: For the distance-based methods (Coreset, KNN), we follow the instruction in [16] to extract the low-dimensional feature and define the distance metric. Specifically, take the output of the last block of the ViT

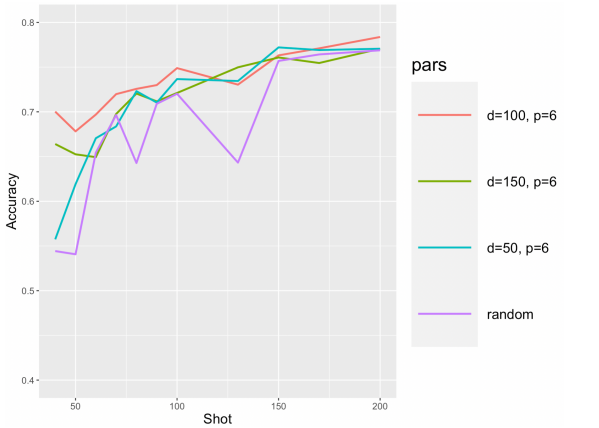

Fig. 10 The image classification accuracy given the budgets (number of training samples) on the CIFAR-10 dataset with various d. Accuracies are plotted versus different budget r (x axis).

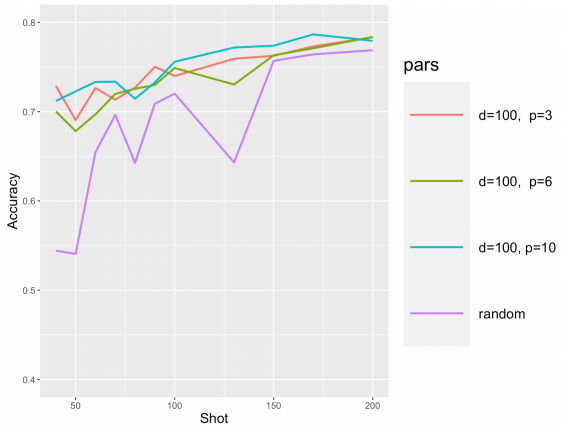

Fig. 11 The image classification accuracy given the budgets (number of training samples) on the CIFAR-10 dataset with various p. Accuracies are plotted versus different budget r (x axis).

model as the image features, and use the l_2 distance as the distance metric. For SPOT, since the properties of space-filling designs are restricted to a relatively low dimension, we further apply a simple Autoencoder [71] with a three-layer Encoder and a three-layer Decoder to reduce the dimension. The principal component analysis is applied when needed.

CNN model: For the distance-based methods (Coreset, KNN), we take the output of the second fully connected layer of the CNN model as the image features, and use the l_2 distance as the distance metric. For SPOT, we use the principal component analysis to reduce the dimension further when needed.

A.4 Parameter Sensitivity

In order to evaluate the robustness of the proposed SPOT algorithm over the parameters in the dimension reduction step, we take the benchmark dataset CIFAR-10 as an example to conduct experiments. Specifically, we test the influence of (1) the number of nodes d for the latent layer in Autoencoder, and (2) the number of principal components p used in PCA.

Specifically, we first fix p to be 6 and vary d among 50, 100, and 150 to explore the influence of d. Results are shown in Fig. 10. We observe that the overall trend of accuracy is upward as the shot size increases for all scenarios. For different d, the increase in accuracy of the proposed SPOT algorithm is stable, while the increase of the random sampling method fluctuates greatly. Moreover, the performance of the proposed SPOT algorithm is stable across different values of d, and outperforms the random sampling algorithm for almost all scenarios.

Then we fix d to be 100 and vary p among 3, 6, and 10 to explore the influence of p. Results are shown in Fig. 11. Similar to the phenomenon in Fig. 10, the overall trend of accuracy is upward as the shot size increases for all scenarios. For different p, the increase in accuracy of the proposed SPOT algorithm is more stable than the random sampling method. Moreover, the performance of the proposed SPOT algorithm has better performance than random sampling in all scenarios and is stable across different values of p.

References of Supplementary Material

[71] Ganggang Dong, Guisheng Liao, Hongwei Liu, and Gangyao Kuang. A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geoscience and Remote Sensing Magazine*, 6(3):44–68, 2018.