Mind the Boundary: Coreset Selection via Reconstructing the Decision Boundary

Shuo Yang[1] Zhe Cao[2] Sheng Guo[3] Ruiheng Zhang[2] Ping Luo[1 4] Shengping Zhang[5] Liqiang Nie[5]

Abstract

Existing paradigms of pushing the state of the art require exponentially more training data in many fields. Coreset selection seeks to mitigate this growing demand by identifying the most efficient subset of training data. In this paper, we delve into geometry-based coreset methods and preliminarily link the geometry of data distribution with models' generalization capability in theoretics. Leveraging these theoretical insights, we propose a novel coreset construction method by selecting training samples to reconstruct the decision boundary of a deep neural network learned on the full dataset. Extensive experiments across various popular benchmarks demonstrate the superiority of our method over multiple competitors. For the first time, our method achieves a 50% data pruning rate on the ImageNet-1K dataset while sacrificing less than 1% in accuracy. Additionally, we showcase and analyze the remarkable cross-architecture transferability of the coresets derived from our approach.

1. Introduction

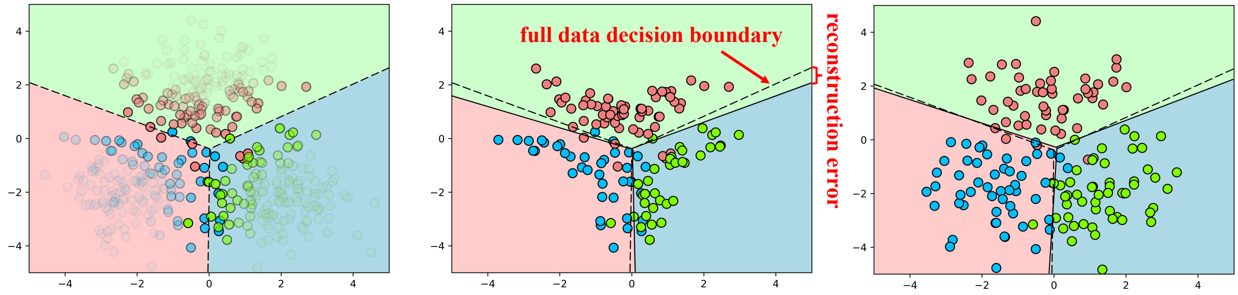
Benefiting from training on datasets of unprecedented scale, language and vision foundational models (Brown et al., 2020; Radford et al., 2021) exhibit the immense potential to actualize Artificial General Intelligence (AGI). However, challenges arise as recent studies reveal a *power law* trend linking the generalization capability of deep neural networks to the volume of their training data, namely, *scaling law* (Hestness et al., 2017; Rosenfeld et al., 2020; Gordon et al., 2021). This trend implies that any further reduction on the testing error may necessitate an order of magnitude more training data, thereby exponentially escalating the computational cost (Kaplan et al., 2020). The consequent computational and storage exhaustion and unsustainable data growth pose significant barriers to the development of general intelligence models, calling for innovative approaches to efficiently utilize training data.

Fortunately, a recent study (Sorscher et al., 2022) provides a potential breakthrough by suggesting that the power-law correlation between the testing error and training data size could be downgraded to exponential scaling. This could be achieved by selecting a high-quality, information-rich subset (*a.k.a*, coreset) from the entire pool of training data. This revelation paves the way for a potential reduction in training costs, while simultaneously retaining performance. Generally, popular coreset selection methods can be stratified into three categories (Guo et al., 2022): geometry-based, score-based, and optimization-based methods. ***Geometry-based*** methods select samples grounded on their geometric characteristics in the feature space, such as distance to class centers (Rebuffi et al., 2017; Castro et al., 2018; Belouadah & Popescu, 2020; Kaddour, 2023), distance to other selected samples (Wolf, 2011; Sener & Savarese, 2018), and distance to feature median (Xia et al., 2022). ***Score-based*** methods rank and choose training samples based on a specific predefined score about model prediction (Maharana et al., 2023; Huang et al., 2023), including the Forgetting score (Toneva et al., 2019), EL2N score (Paul et al., 2021), and uncertainty score (Liu et al., 2019; He et al., 2023). However, most of these approaches select data based on heuristicly designed metrics and are devoid of guaranteed generalization ability for the coreset, making it hard to justify the effectiveness and theoretical property of the coreset. To overcome this, ***optimization-based*** methods (Tukan et al., 2023; Tan et al., 2023; Abbas et al., 2023) strive to bestow superior theoretical properties on the coreset by optimizing the selected data to ensure they mirror a similar gradient direction (Killamsetty et al., 2021a; Mirzasoleiman et al., 2020), influence function (Koh & Liang, 2017; Yang et al., 2022; Pooladzandi et al., 2022), or validation error (Killamsetty et al., 2021b; Borsos et al., 2020) to the full training data. Though equipped with solid theoretical underpinnings, these methods grapple with severe challenges when scaling to large-scale datasets due to the intricate bilevel optimization.

Though various geometry-based methods have been proposed in the literature (Sener & Savarese, 2017; Cohen-

[1]The University of Hong Kong, Hong Kong [2]Beijing Institute of Technology, China [3]MYBank, Ant Group, China [4]Shanghai AI Lab, China [5]Harbin Institute of Technology, China. Correspondence to: Shengping Zhang <s.zhang@hit.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Full training data (b) BoundarySet (c) BoundarySet-CCS

Figure 1. The t-SNE visualization of the feature space of three randomly chosen classes, each with 150 samples from the CIFAR-10. For each data point, we compute its distance d to its nearest decision boundary using the Equation 1. (a) All data points are plotted, samples with a darker color indicate a smaller calculated distance d. *It can be observed that our computed distance is an accurate estimation.* (b) BoundarySet selects 30% data points with smallest d. The dashed line represents the old decision boundary (as in Figure. (a)), and the solid line represents the new decision boundary learned on the BoundarySet. The area between these two boundaries is defined as the reconstruction error. (c) BoundarySet-CCS collects data points closer to the decision boundary while simultaneously guaranteeing the distribution coverage.

Addad et al., 2021; Huang et al., 2019), most, if not all, of them are heuristics-based, without any guarantee of the generalization capability of models. In this paper, we undertake the pioneering effort to bridge the geometric properties of selected coresets and the generalization capabilities of models trained on them. The key idea behind our method is to select a sub-trainset to reconstruct the decision boundary of the model learned on the full-trainset. Specifically, we first introduce a concept of *decision boundary reconstruction error*, which refers to the quantifiable discrepancy between the decision boundaries of two models trained with the full data and selected subset, respectively. We prove that the decision boundary reconstruction error can act as a strict upper bound for the generalization error gap of these two models. From our theoretical findings, we deduce a significant conclusion: *a coreset with a low decision boundary reconstruction error will enable a model trained on it to exhibit generalization abilities closely mirroring those of a model trained on the full dataset.* Inspired by the crucial role of *support vectors* (Cortes & Vapnik, 1995) in shaping the decision boundary, we propose a method to collect samples near the decision boundaries of deep neural networks to construct a coreset that minimizes the *decision boundary reconstruction error*. Our empirical studies demonstrate that the support vectors chosen by different deep neural networks on the same dataset have a considerable overlap. This finding validates the cross-model transferability of our decision boundary-based coreset selection method. Our method achieves state-of-the-art performances on the CIFAR-10/100 (Krizhevsky et al., 2009) and ImageNet-1K (Russakovsky et al., 2015) datasets, and for the first time, achieves less than 1% accuracy drop when pruning over 50% training samples on the challenging ImageNet-1K dataset.

Before delving into details, we summarize our contributions as below:

- Motivated by the association between a model's decision boundary and its generalization capability (Mickisch et al., 2020; Lei et al., 2022), our work innovatively proposes to construct a coreset aimed at reconstructing the decision boundary shaped by the full data. Moreover, we have developed a theoretical framework that clearly delineates the relationship between the error in decision boundary reconstruction and the discrepancy in generalization, providing robust theoretical support for our boundary-based coreset selection method.

- To identify the *support vectors* of a deep neural network, we propose a novel perturbation-based method to approximate the distance of a data point to its nearest decision boundary. Additionally, we employ a coverage-centric data sampling strategy, which selects samples closer to the decision boundary while simultaneously ensuring distribution coverage. Both components are novel, well-motivated, and crucial for achieving superior performance.

- Our proposed method demonstrates favorable performance across several challenging benchmarks, consistently achieving state-of-the-art results in all datasets and problem settings. Specifically, on the ImageNet-1K dataset, it attains a 30% lossless pruning ratio and exhibits less than a 1% accuracy drop when half of the training data is removed. Furthermore, we have conducted exhaustive analysis on the decision boundary reconstruction error and the cross-architecture transferability of our method. This thorough analysis provides the community with valuable insights into the mechanics of decision boundary-based coreset selection.

2. Motivation and Theoretical Analysis

An optimal coreset should ideally exhibit comparable generalization performance to the full data when deployed for model training. However, given that the test distribution is untouchable, it proves computationally challenging to directly optimize the coreset to minimize the generalization error (Borsos et al., 2020). In light of the intrinsic link between a discriminative model's decision boundary and its generalization performance (Li et al., 2018; Mickisch et al., 2020; Lei et al., 2022), our work introduces a novel perspective to construct a coreset with a focus on decision boundary reconstruction.

In support of our motivation, two principal questions must be elucidated: *(1) how to quantitatively measure a selected coreset's proficiency at reconstructing the model's decision boundary*, and *(2) what is the relationship between the decision boundary reconstruction error and the model's generalization gap*. In this section, we endeavor to shed light on these questions. We first provide a formal definition of the decision boundary reconstruction error, quantifying the divergence between the class boundaries of two models learned on the full data and the selected coreset respectively. Then, we theoretically prove that the generalization gap introduced by the coreset can be effectively bounded by the decision boundary reconstruction error.

Given a training set, $\mathcal{S} = (x_i, y_i)_{i=1}^m$, where $x_i \in \mathbb{R}^z$ denotes input data in z-dimensional real space, and $y_i \in [c] = \{1, \ldots, c\}$ refers to the corresponding class labels, with c being the total number of classes. The size of the training sample is $m = |\mathcal{S}|$. We assume that all data (x_i, y_i) are independent and identically distributed (*i.i.d.*) random variables drawn from a certain data distribution $\mathcal{D}$. We represent the classifier as $f_\theta(x) : \mathbb{R}^z \to \mathbb{R}^c$, which is essentially a neural network parameterized by θ. The output, $f_\theta(x)$, is assumed to be a c-dimensional vector that acts as a discrete probability density function. Here, $f_\theta^{(i)}(x)$ denotes the i-th component of $f_\theta(x)$, such that $\Sigma_{i=1}^c f_\theta^{(i)}(x) = 1$. We then define the function $T(f_\theta, x) = \{i \in \{1, \cdots, c\} | f_\theta^{(i)}(x) = \max_j f_\theta^{(j)}(x)\}$, which represents *the set of predicted labels* by f_θ for a given input x. Denote $\mathcal{S}'_\eta = (x_i, y_i)_{i=1}^n$ is a η-coreset of $\mathcal{S}$, $n < m$, $\frac{n}{m} = \eta$, $\mathcal{S}'_\eta \subset \mathcal{S}$, we define the decision boundary reconstruction error of η-coreset $\mathcal{S}'_\eta$ to $\mathcal{S}$ over $f_\theta(x)$ as below.

Definition 1 (Decision Boundary Reconstruction Error of η-coreset)**.** *Let $f_\theta(x) : \mathbb{R}^n \to \mathbb{R}^c$ be a neural network for classification parameterized by θ, where $\theta \sim \mathcal{A}(\mathcal{S})$ is returned by leveraging the learning algorithm $\mathcal{A}$ on the training set $\mathcal{S}$, which is sampled from the data generating distribution $\mathcal{D}$. Denote $\mathcal{S}'_\eta \subset \mathcal{S}$ as a η-coreset of $\mathcal{S}$, where $\frac{|\mathcal{S}'_\eta|}{|\mathcal{S}|} = \eta$. Then, we say $\mathcal{S}'_\eta$ has a decision boundary reconstruction error of ϵ to $\mathcal{S}$ over $f_\theta(x)$ if*

$$\mathbb{E}_{\mathcal{D}} \mathbb{E}_{\theta \sim \mathcal{A}(\mathcal{S}), \theta' \sim \mathcal{A}(\mathcal{S}'_\eta)} \left[\mathbb{I}\left(T(f_\theta, x) \neq T(f_{\theta'}, x) \right) \right] = \epsilon,$$

where $\mathbb{I}(\cdot)$ is the indicator function. It is noteworthy that *$\mathcal{D}$ is the data generation distribution, not the training data distribution.* If a selected coreset $\mathcal{S}'_\eta$ possesses a *zero* decision boundary reconstruction error in relation to $\mathcal{S}$ with respect to $f_\theta(x)$, it means that two classifiers, f_θ and $f_{\theta'}$, each trained on the full data $\mathcal{S}$ and the coreset $\mathcal{S}'_\eta$, will yield identical category predictions for all samples x drawn from the data generation distribution $\mathcal{D}$. Denote the expected risk of the model trained on the dataset $\mathcal{S}$ over the distribution $\mathcal{D}$ as

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{E}_{\theta \sim \mathcal{A}(\mathcal{S})} \left[\mathbb{I}\left(y \notin T\left(f_\theta, x \right) \right) \right]$$

We have the following theorem.

Theorem 1. *If the selected η-coreset $\mathcal{S}'_\eta$ has a decision boundary reconstruction error of ϵ to $\mathcal{S}$ over $f_\theta(x)$, then we have*

$$\left| \mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) - \mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S}'_\eta)) \right| \leq \epsilon.$$

Theorem 1 establishes a relationship wherein the disparity between the expected risks of $\mathcal{A}(\mathcal{S})$ and $\mathcal{A}(\mathcal{S}'_\eta)$ can be bounded by the differences in their respective decision boundaries. It can be easily proved as follows.

Proof.

$$\begin{aligned}
&\left| \mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) - \mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S}'_\eta)) \right| \\
&= \left| \mathbb{E}_{\mathcal{D}, \mathcal{A}(\mathcal{S})} \left[\mathbb{I}\left(y \notin T\left(f_{\boldsymbol{\theta}}, \mathbf{x} \right) \right) \right] - \mathbb{E}_{\mathcal{D}, \mathcal{A}\left(\mathcal{S}'_\eta\right)} \left[\mathbb{I}\left(y \notin T\left(f_{\boldsymbol{\theta}}, \mathbf{x} \right) \right) \right] \right| \\
&= \left| \mathbb{E}_{\mathcal{D}, \mathcal{A}(\mathcal{S}), \mathcal{A}\left(\mathcal{S}'_\eta\right)} \left[\mathbb{I}\left(y \notin T\left(f_{\boldsymbol{\theta}}, \mathbf{x} \right) \right) - \mathbb{I}\left(y \notin T\left(f_{\boldsymbol{\theta}'}, \mathbf{x} \right) \right) \right] \right| \\
&\leq \mathbb{E}_{\mathcal{D}, \mathcal{A}(\mathcal{S}), \mathcal{A}\left(\mathcal{S}'_\eta\right)} \left[\left| \mathbb{I}\left(y \notin T\left(f_{\boldsymbol{\theta}}, \mathbf{x} \right) \right) - \mathbb{I}\left(y \notin T\left(f_{\boldsymbol{\theta}'}, \mathbf{x} \right) \right) \right| \right] \\
&\leq \mathbb{E}_{\mathcal{D}, \mathcal{A}(\mathcal{S}), \mathcal{A}\left(\mathcal{S}'_\eta\right)} \left[\mathbb{I}\left(T\left(f_{\boldsymbol{\theta}}, \mathbf{x} \right) \neq T\left(f_{\boldsymbol{\theta}'}, \mathbf{x} \right) \right) \right] = \epsilon
\end{aligned}$$

Theorem 1 indicates that a coreset $\mathcal{S}'_\eta$, with a lower decision boundary reconstruction error ϵ, will result in a smaller generalization gap, thus leading to performance closer to the full data $\mathcal{S}$. This directly motivates us to construct coresets that contribute most to the shaping of the decision boundary.

3. Methodology

Motivated by Theorem 1, this section is dedicated to proposing a methodology for constructing a coreset capable of reconstructing the decision boundaries of a model trained with full data. In Section 3.1, we introduce a novel approach designed to estimate the distance of each sample from its nearest decision boundary. Building on this, Section 3.2 introduces two boundary-based coreset selection strategies, including a distance-based sampling and a coverage-based sampling strategy.

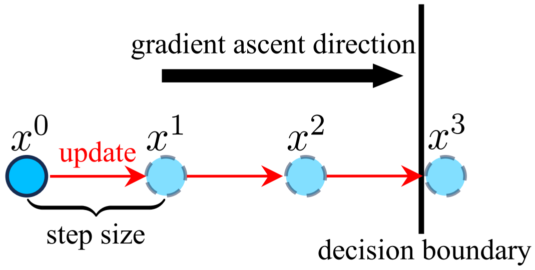


Figure 2. Illustration of the distance approximation process. At each step k, we update the data x^k towards the gradient ascent direction. The minimal step that a data x required for crossing the decision boundary is considered as an approximation of its distance to the decision boundary.

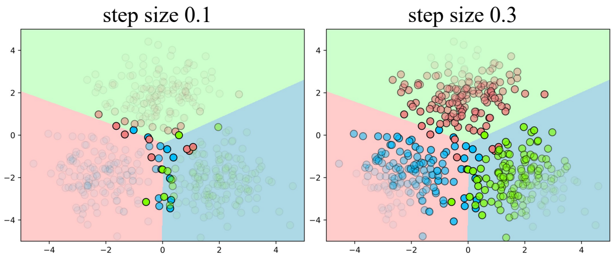


Figure 3. The effect of step size α in Equation 1. The data points are visualized with different levels of transparency, indicating their calculated distance d. As illustrated in the figure, a smaller step size is more effective in accurately identifying samples in close proximity to the decision boundary.

3.1. Distance to Decision Boundary

Drawing inspiration from the principles of Support Vector Machines (SVMs) (Cortes & Vapnik, 1995), it is observed that a minimal set of data points, situated in proximity to the decision boundary (termed Support Vectors), can facilitate the reconstruction of a comparable or identical decision boundary. The challenge in the context of deep networks, however, lies in the non-trivial task of identifying these support vectors, given that their decision boundary frequently exhibits high complexity. One could potentially estimate the distance to the decision boundary by examining the minimal distance between instances belonging to separate classes (Ducoffe & Precioso, 2018), but such an approach often yields coarse evaluations and incurs significant computational overhead.

Indeed, adversarial attack (Chakraborty et al., 2018; Kurakin et al., 2017) – which aims to introduce the smallest possible perturbation to a data point, thereby prompting it to cross the decision boundary – provides us a possible way to approximate the distance to the decision boundary. Specifically, we seek to estimate the minimum distance from a training sample to the decision boundary by counting the minimal iteration step k that the Projected Gradient Descent (PGD) algorithm (Madry et al., 2018) necessitates in order to produce its misclassified adversarial counterpart. This approach aligns with the techniques employed in various adversarial training methods (Zhang et al., 2019; Ding et al., 2020; Zhang et al., 2021).

Given a training sample (x, y), PGD works as follows:

$$x^{(k+1)} = \Pi_{\mathcal{B}[x^{(0)}]}\left(x^{(k)} + \alpha\,\text{sign}(\nabla_{x^{(k)}}\ell(f_\theta(x^{(k)}), y))\right) \tag{1}$$

where $x^{(0)}$ is the starting point initialized by x, $x^{(k)}$ is adversarial data at step k, α is the step size, ℓ is the loss function, and $\Pi_{\mathcal{B}[x^{(0)}]}[\cdot]$ is the projection function that projects the adversarial data back into the ϵ-ball centered at $x^{(0)}$ (not necessary in our method). This process iterates until the adversarial data $x^{(k)}$ has crossed the decision boundary (*i.e.*, $f_\theta(x^{(k)}) \neq y$) or the number of steps k reaches the maximally allowd iteration number K (*i.e.*, $k = K$).

In each step, PGD updates the adversarial instance along the direction that maximizes the alteration in the model's prediction. Therefore, the minimum iteration number k, required by the PGD method to generate an adversarial variant $\tilde{x}$ for a given data point (x, y) to cross the decision boundary, can serve as an approximation of the smallest distance from a data point to the decision boundary, represented as $d(x, y) = k, k \in [0, K]$. Figure 2 depicts the process of distance approximation in Equation 1. Figure 1 (a) visualizes all data points with varying degrees of transparency corresponding to their computed distance d. This figure demonstrates that our computed d is an accurate estimation of their ground-truth distance to the decision boundary. Additionally, Figure 3 explores the impact of step size α as formulated in Equation 1. Employing a smaller step size allows for minor perturbations of the instance x in each iteration and facilitates a more nuanced differentiation between samples across different distances. It is particularly effective in disentangling samples that are closely situated, thereby enhancing the granularity of our analysis.

3.2. Boundary-based Coreset Selection

Distance-based Sampling. Based on the distance computed in Section 3.1, we can naturally collect samples that are closest to the decision boundary to form the coreset, termed BoundarySet (as depicted in Figure 1 (b)). The BoundarySet exhibits favorable performance, especially at lower pruning ratios, achieving 50% lossless pruning on CIFAR-10 and 30% on both CIFAR-100 and ImageNet-1K, as shown in Figure 4. This success is attributed to the precise reconstruction of the decision boundary by the selected samples. However, a sharp decrease in performance is observed with increased pruning ratios, reflecting the challenges in maintaining decision boundary with sparser data. As the selection ratio decreases, the data distribution becomes less representative of the original set, thereby affecting the boundary reconstruction. *In essence, at higher selection ratios, the decision boundary reconstruction is primarily influenced by those samples that are located near the boundary. Conversely, at lower selection ratios, the reconstruction's fidelity hinges on the ability to recover the overall data distribution.* Thus, while BoundarySet excels in environments with less pruning, its effectiveness diminishes as the selection ratio lowers, underscoring the importance of a balanced dataset for accurate decision boundary delineation.

Coverage-Centric Sampling. To counteract the aforementioned problem, we incorporate the coverage-centric sampling (CCS) strategy (Zheng et al., 2022), which ensures both diversity and representativeness in the selected coreset when the selection ratio is extremely low. Specifically, all training samples are divided into $K + 1$ non-overlapping groups based on their distance to the boundary, $d(x, y) \in [0, K]$. An initial sample selection budget is uniformly allocated across these groups, contingent upon the desired selection rate. If any group contains fewer samples than its allocated budget, the surplus budget is evenly redistributed among the remaining groups. The efficacy of the coverage-centric sampling is illustrated in Figure 1(c). Compared to the distance-based selection BoundarySet shown in Figure 1(b), the BoundarySet-CCS not only selects samples nearer to the decision boundary but also ensures a comprehensive distribution coverage. The effectiveness of both strategies is empirically validated in our experiments. Detailed methodology of the coverage-centric sampling algorithm can be found in the related literature (Zheng et al., 2022). The BoundarySet-CCS selection process is shown in Algorithm 1.

Algorithm 1 Boundary-aware coreset selection.

```
Input: Full dataset $\mathcal{S} = (x_i, y_i)_{i=1}^m$.
Required: Selection ratio $\eta$; step size $\alpha$; max step $K$; loss
  function $\ell$; classification model $f(\theta)$ trained on $\mathcal{S}$.
// Section. 3.1: Estimating Distance-to-Boundary for Each
  Instance
for $i = 1, \dots, m$ do
    $x^{(0)} \leftarrow x_i$
    for $k = 0, \dots, K$ do
        $d(x_i, y_i) \leftarrow k$
        if $\arg\max_i f_\theta(x^{(k)}) = y_i$ then
            $x^{(k+1)} \leftarrow x^{(k)} + \alpha\,\mathrm{sign}(\nabla_{x^{(k)}} \ell(f_\theta(x^{(k)}), y_i))$
        else
            Break
// Section. 3.2: Coverage-Centric Sampling
$q \leftarrow m \times \eta$;
$\mathcal{S}'_\eta \leftarrow \varnothing$;
$\mathcal{D} \leftarrow \{\mathbb{D}_i, : \mathbb{D}_i$ consists of examples with distance $i, i =
  0, \dots, k\}$
while $\mathcal{D} \neq \varnothing$ do
    $\mathbb{D}_{\mathrm{min}} \leftarrow \arg\min_{\mathbb{D} \in \mathcal{D}} |\mathbb{D}|$
    $m_D \leftarrow \min\{|\mathbb{D}_{min}|, \lfloor \frac{q}{|\mathcal{D}|} \rfloor\}$
    $\mathcal{S}_D \leftarrow$Randomly sample $m_D$ examples from $\mathbb{D}_{\mathrm{min}}$
    $\mathcal{S}'_\eta \leftarrow \mathcal{S}'_\eta \cup \mathcal{S}_D$;
    $\mathcal{D} \leftarrow \mathcal{D} \backslash \mathbb{D}_{min}$
    $m \leftarrow m - m_D$
Output: The $\eta$-coreset $\mathcal{S}'_\eta = (x_i, y_i)_{i=1}^n$, $n < m$.
```

4. Experiments

4.1. Experimental Setup and Implementation Details

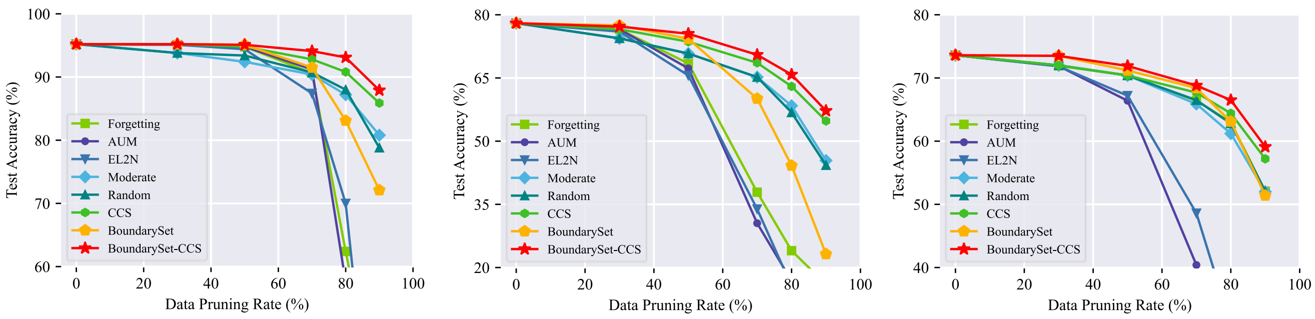
Datasets and Training Details. We evaluate the effectiveness of our method on three popularly used datasets, i.e., CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-1K (Russakovsky et al., 2015). For the CIFAR-10 and CIFAR-100, we use ResNet-18 (He et al., 2016) as the network architecture. For all coresets with all pruning rate levels, we train models for 40,000 iterations with a 256 batch size. We use the SGD optimizer (0.9 momentum and 0.0002 weight decay) with a 0.1 initial learning rate. The learning rate scheduler is the cosine annealing learning rate scheduler with a 0.0001 minimum learning rate. We use a 4-pixel padding crop and a randomly horizontal flip as data augmentation. For the ImageNet-1K, We use ResNet-34 (He et al., 2016) as the network architecture. For all coresets with different pruning rates, we train models for 300,000 iterations with a 256 batch size. We use the SGD optimizer (0.9 momentum and 0.0001 weight decay) with a 0.1 initial learning rate. The learning rate scheduler is the cosine annealing learning rate scheduler. For the distance estimation process, we employ a step size of 0.002 and max step of 10 for CIFAR-10, a step size of 0.001 and max step of 20 for CIFAR-100, and a step size of 0.0001 and max step of 50 for ImageNet-1K.

Competitors. We compare our method BoundarySet and the enhanced version – BoundarySet-CSS with the following popular methods:

(i) *Random.* Samples are randomly selected from the training set to form a coreset.

(ii) *Forgetting* (Toneva et al., 2019). The process where a training sample transitions from being correctly classified to being incorrectly classified is termed a 'forgetting' event, which repeatedly occurs during training. The forgetting score measures the frequency of a sample being forgotten throughout the entire course of training. Coreset selection strategies based on forget scores involve selecting samples with high forgetting scores.

(iii) *EL2N* (Paul et al., 2021). The L2 norm of the error



(a) Evaluation on CIFAR-10.

(b) Evaluation on CIFAR-100.

(c) Evaluation on ImageNet-1k.

Figure 4. Comparative performance of proposed method against other baselines across CIFAR-10, CIFAR-100 and ImageNet datasets at various selection ratios. In every evaluated scenario, BoundarySet-CCS consistently exhibits performance that surpasses all established baselines. Notably, BoundarySet also shows commendable effectiveness at low pruning rates. The performance degradation observed at higher pruning rates can be attributed to a decrease in data distribution coverage (Zheng et al., 2022). The CCS variant of BoundarySet has been optimized to address this specific issue, resulting in superior performance across the entire spectrum of pruning rates.

vector of a training sample is referred to as the EL2N score. Coreset selection strategies based on the EL2N score aim to retain samples with large errors.

(iv) *Moderate* (Xia et al., 2022). It proposes a distance-based score for one-shot coreset selection. The Moderate selection method considers samples close to the median value of the feature space as more important.

(v) *AUM*. (Pleiss et al., 2020) Area Under the Margin (AUM) statistic leverages the distinct training behaviors of correctly labeled and mislabeled samples to identify samples that should be collected into the coreset. By adding a specially designed class with deliberately mislabeled samples, it sets an upper AUM limit to effectively detect mislabeled data.

(vi) *CCS*. Coverage-centric Coreset Selection (CCS) (Zheng et al., 2022) is an innovative data selection method, distinguished by its emphasis on maintaining coverage in high-density areas of a dataset. Unlike state-of-the-art methods that prioritize difficult samples and prune easy data, CCS focuses on maintaining comprehensive data coverage using a stratified sampling strategy.

4.2. Comparison to the State of the Art

We compare our method against six popular methods across three datasets, demonstrating the superior performance of our approach under varying data pruning rates. As depicted in Figure 4, across all datasets, our method outperforms all SOTA methods at lower pruning rates. This is attributed to our method's precise selection of samples near the decision boundary, which are crucial for accurate reconstruction of the decision boundary.

CIFAR-10. As depicted in Figure 4 (a), our method achieves impressive performance on the CIFAR-10 dataset, with zero performance decrease at a 50% pruning rate for both BoundarySet and BoundarySet-CCS. Even at 90% pruning rate, the BoundarySet-CCS still maintains 87.9% accuracy, which remarkably outperforms all competitors.

CIFAR-100. With a pruning rate of 30% on CIFAR-100 dataset, the accuracy of BoundarySet is 77.48%, which represents only a 0.52% decrease compared to the whole-dataset training, i.e., without pruning, and a performance increase of nearly 1% over CCS. When the pruning rate is increased to 50%, BoundarySet-CCS achieves an accuracy of 75.5%, representing a 2.5% reduction from the no-pruning case.

ImageNet-1K. In the case of ImageNet-1K, a 30% pruning rate yielded an accuracy of 73.5%, representing a minimal decline of only 0.1% compared to the no-pruning scenario. Moreover, delving deeper into ImageNet-1K with increased pruning rates of 70%, 80%, and 90%, the BoundarySet-CCS method exhibits accuracies of 68.8%, 66.5%, and 59.1%, respectively. In these cases, the integration of CCS plays a pivotal role, contributing to incremental improvements in the accuracy of 0.4%, 2.6%, and 7.7% at each respective pruning rate.

With increasing pruning rates, the decline of testing accuracy is observed not only in BoundarySet but also in Forgetting, EL2N, and AUM. We concur with the perspective offered by CCS (Zheng et al., 2022): a limited range of data distribution coverage directly results in lower testing accuracy. At high pruning rates, the data selection of BoundarySet at the decision boundary inherently leads to biased sampling and reduced data distribution coverage. In contrast, Moderate, CCS, and Random focus on data distribution coverage, hence maintaining high testing accuracy even at elevated pruning rates. In light of this, our enhanced method

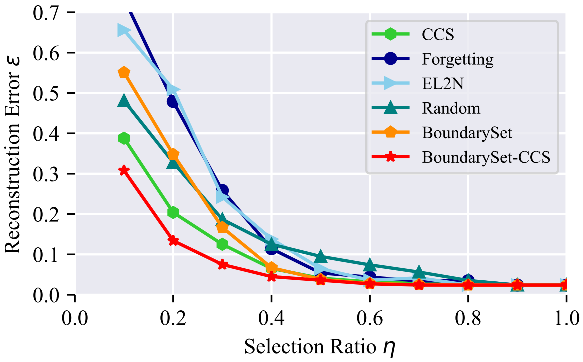


Figure 5. Curve of the decision boundary reconstruction error ϵ at various selection ratios η on CIFAR-100. The trend corresponding to Figure 4 corroborates the validity of Theorem 1, and demonstrates that the decision boundary reconstruction error ϵ serves as a reliable metric for the quality of the selected coreset.

BoundarySet-CSS integrates consideration of both decision boundary proximity (i.e., the data difficulty of coreset) and data distribution coverage (i.e., the data distribution diversity of coreset), thereby outperforming all baselines across all datasets and pruning rates.

4.3. Analysis of the Reconstruction Error

According to the Definition 1, we can estimate the decision boundary reconstruction error ϵ of a given η-coreset, by evaluating the prediction discrepancy of the two models trained on the full data $\mathcal{S}$ and the coreset $\mathcal{S}'_\eta$ on the data generating distribution $\mathcal{D}$. To simulate the data generating distribution $\mathcal{D}$, we trained a conditional BigGAN (Zhao et al., 2020) to synthesize 100,000 fake images for the CIFAR-10 dataset. The synthetic images are used for calculating the ϵ-η curve for all coreset selection methods on CIFAR-10 in Figure 5.

As illustrated in Figure 5, with the increase in the selection ratio η, the reconstruction error ϵ for all methods gradually decreases, aligning with the description in Theorem 1. Given that the sum of the selection ratio η and pruning rate equals 1, Figure 5 can be mirrored and vertically flipped to compare with Figure 4. This comparison reveals a strict consistency between the declining curve of the reconstruction error ϵ and the trend of the test accuracy curve. This strongly validates the accuracy of Theorem 1, demonstrating that the reconstruction error ϵ serves as a reliable metric for evaluating the quality of a coreset.

At higher selection ratios η, the reconstruction error ϵ focuses on the difficulty of data selected by the coreset, namely, its capacity to delineate decision boundaries accurately. ***Conversely, at lower selection ratios η, the reconstruction error ϵ is more concerned with the diversity of the selected data, that is, whether the coreset's data distribution can represent the true data distribution.*** Thus, the BoundarySet achieves lower errors at high selection ratios η, while can not maintain low reconstruction errors at low ratios. Our enhanced model, BoundarySet-CCS, not only identifies the most useful support vectors for reconstructing decision boundaries, like the BoundarySet, but also achieves better distribution coverage. Both the two components are essential for achieving the low reconstruction error. Therefore, as seen in Figure 5, our method surpasses all baselines, exhibiting the smallest reconstruction error ϵ at any selection ratios η. It validates that our method possesses excellent decision boundary reconstruction capability.

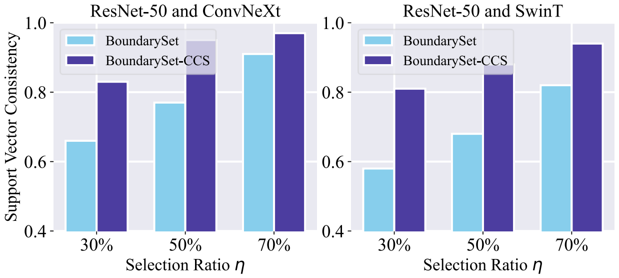


Figure 6. Cross-architecture transferability of the support vectors. We compare the consistency of support vectors (coresets) selected by BoundarySet and BoundarySet-CCS using ResNet-50, ConvNeXt, and SwinT at different selection ratios η, where the consistency is defined as the intersection ratio of coresets.

4.4. Cross-Architecture Transferability of the Support Vectors in Deep Neural Networks

Despite the excellent results achieved by our proposed method, we remain focused on an essential question: *Is the selected coreset universally applicable to different models?* As shown in Figure 6, we compared the consistency of the selected coreset using different backbone networks, namely, ResNet-50, ConvNeXt (Liu et al., 2022), and Swin-transformer (Liu et al., 2021). Support Vector Consistency, indicating the intersection ratio of the coresets selected by different networks, suggests that higher consistency corresponds to better generalizability of the coreset method. It is observed that the consistency of our method consistently exceeds 0.5 under any circumstances, and it increases as the selection rate rises. This indicates that the coresets selected by different model architectures are generally similar, confirming the transferability of our method.

The integration of CCS into BoundarySet, on average, improves consistency by about 18%. The underlying reason for this improvement is that CCS places greater emphasis on the representativeness of data distribution, which is independent of the network architecture employed by the model. However, consistency across different architecture families presents a greater challenge to the method's transferability. Both ResNet-50 and ConvNeXt belong to the CNN fam-

Selection → Test Model	ImageNet-1K Acc. (%)
ResNet-50 → SwinT	77.03
ConvNeXt → SwinT	77.64
Swin No Pruning	78.46
ResNet-50 → ConvNeXt	76.91
SwinT → ConvNeXt	76.52
ConvNeXt No Pruning	77.88

Table 1. The cross-architecture performance of BoundarySet-CCS at a 50% selection ratio on ImageNet-1K. 'ResNet-50 → SwinT' denotes the classification training of SwinT using a coreset selected by ResNet-50. Both intra-family and cross-family transfer training exhibit superior performance. Notably, our method achieves a remarkable feat by exhibiting a less than 1% performance decrease at a 50% selection ratio on the ImageNet-1K dataset.

ily, whereas Swin-Transformer is part of the ViT family. The consistency between ResNet-50 and ConvNeXt is always marginally higher than that between ResNet-50 and Swin-Transformer, indicating a smaller gap within the same family and lesser transferability between different families. More detailed comparisons, as shown in Table 1, reveal that using a coreset obtained with ResNet-50 to train ConvNeXt results in less than a 1% drop in testing accuracy, while using a coreset collected with Swin-Transformer for training ConvNeXt leads to a 1.36% decrease due to cross-family differences.

In conclusion, the aforementioned experiments demonstrate that our method exhibits excellent transferability, even in cross-family scenarios. It is noteworthy to mention an additional point: on ImageNet-1k, with a pruning rate of 50%, our method is the first to achieve a reduction in testing accuracy within 1%.

5. Discussion with Other Boundary-aware Methods

Recent works in dataset pruning highlight that samples near the decision boundary contribute more to model performance (Sorscher et al., 2022; Liu et al., 2019; Li et al., 2023). However, due to the high complexity of deep neural networks' decision boundaries, it is non-trivial to directly locate those samples close to the boundary. Chen et al. (2021) emphasizes the importance of decision boundaries but ultimately chooses to use the distance to the clustering center as an estimation of the distance to the decision boundary. Ducoffe & Precioso (2018) estimate the distance to the decision boundary using samples with low classification confidence. This approach, though innovative, lacks comprehensive theoretical analysis and oversimplifies the relationship between low classification confidence and proximity to the decision boundary. Similarly, Liu et al. (2019) use model-predicted uncertainty as a proxy. Yet, prediction uncertainty is widely acknowledged as misleading (Abdar et al., 2021) and will inevitably lead to performance degradation of networks trained with the coreset. Furthermore, there is a lack of analysis on how well the selected samples reconstruct the decision boundary and the correlation between decision boundary reconstruction error and model generalization error. This paper pioneers a coreset selection method aiming directly at decision boundary reconstruction, establishing a theoretical link between decision boundary reconstruction error and model generalization ability, contributing innovatively to this research field.

6. Limitation

Although exceptional performances on multiple datasets have been achieved, like many contemporary methods, BoundarySet is based on the assumption of a good feature representation space, which is difficult to obtain on sparse data. In the case of high pruning rates, the small number of samples and biased data distribution pose a tremendous challenge to the training of the feature extractor. This is one of the main reasons for the subpar performance of contemporary methods at high pruning rates (Guo et al., 2022). To relieve this problem, we developed an enhanced version, BoundarySet-CCS, which addresses the performance issues encountered at high pruning ratios. However, the detailed analysis of the intricate interplay between the sample selection process and feature extractors, and the identification of samples that can enhance training efficiency, remains a challenge that requires future exploration.

7. Conclusion

This paper discovered that samples positioned near the decision boundary play an important role in reconstructing the decision boundary in dataset pruning. In response to this discovery, we created a decision-boundary-based coreset selection method. Theoretically, we developed a theoretical framework to clarify the relationship between decision boundary error and generalization discrepancy, providing strong theoretical support for our method. To provide robust theoretical support, we developed a theory framework delineating the relationship between decision boundary error and generalization discrepancy. Particularly, we presented a novel perturbation-based technique for estimating the distance of data points from the decision boundary. In addition, to increase the variety of the coreset at high pruning rates, we implemented a coverage-focused data sampling strategy. Extensive experiments on three datasets revealed that our method outperforms competing baseline models significantly. Ablation experiments on several backbone networks demonstrated the cross-architecture transferability of the proposed method.

Impact Statement

This paper presents a coreset selection algorithm which is generally for efficient deep learning. Our work makes significant contributions to dataset compression, data selection, algorithmic efficiency enhancement, and data privacy protection. The algorithm ensures that the coreset, even when highly pruned, still enables the model to reconstruct decision boundaries precisely. Given the ever-increasing size of datasets and the advent of the era of large models, our research offers a practical and viable technological solution for enhancing computational resource utilization. While there may be many other potential impacts, we believe that our approach does not entail any negative ethical or moral implications.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 62272134 and 62236003), and the Shenzhen College Stability Support Plan (No. GXWD20220817144428005). We thank Bo Zhao from BAAI for his invaluable contributions during the revision process of this paper.

References

Abbas, A. K. M., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Aljundi, R., Lin, M., Goujaud, B., and Bengio, Y. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Belouadah, E. and Popescu, A. Scail: Classifier weights scaling for class incremental learning. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020.

Borsos, Z., Mutny, M., and Krause, A. Coresets via bilevel optimization for continual learning and streaming. *Advances in Neural Information Processing Systems*, 33: 14879–14890, 2020.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., Schmid, C., and Alahari, K. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Chen, Y., Huang, J., Zhu, J., Zhu, Z., Yang, T., Huang, G., and Du, D. Face-nms: A core-set selection approach for efficient face recognition. *arXiv preprint arXiv:2109.04698*, 2021.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 169–182, 2021.

Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J., and Zaharia, M. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Ding, G. W., Sharma, Y., Lui, K. Y. C., and Huang, R. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkeryxBtPB.

Ducoffe, M. and Precioso, F. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020.

Gordon, M. A., Duh, K., and Kaplan, J. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, 2021.

Guo, C., Zhao, B., and Bai, Y. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pp. 181–195. Springer, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

He, M., Yang, S., Huang, T., and Zhao, B. Large-scale dataset pruning with dynamic uncertainty. *arXiv preprint arXiv:2306.05175*, 2023.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Huang, L., Jiang, S., and Vishnoi, N. Coresets for clustering with fairness constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Huang, X., Liu, Z., Liu, S.-Y., and Cheng, K.-T. Efficient quantization-aware training with adaptive coreset selection. *arXiv preprint arXiv:2306.07215*, 2023.

Kaddour, J. The minipile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*, 2023.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Khan, M. A., Hamila, R., and Menouar, H. Clip: Train faster with less data. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 34–39, 2023. doi: 10.1109/BigComp57234.2023.00014.

Killamsetty, K., Durga, S., Ramakrishnan, G., De, A., and Iyer, R. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021a.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., and Iyer, R. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021b.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research). 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJm4T4Kgx.

Lei, S., He, F., Yuan, Y., and Tao, D. Understanding deep learning via decision boundary. *arXiv preprint arXiv:2206.01515*, 2022.

Li, K., Persaud, D., Choudhary, K., DeCost, B., Greenwood, M., and Hattrick-Simpers, J. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nature Communications*, 14(1): 7283, 2023.

Li, Y., Ding, L., and Gao, X. On the decision boundary of deep neural networks. *arXiv preprint arXiv:1808.05385*, 2018.

Liu, K., Liu, W., Cheng, J., and Lu, X. Uhrp: Uncertainty-based pruning method for anonymized data linear regression. In *Database Systems for Advanced Applications: DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22–25, 2019, Proceedings 24*, pp. 19–33. Springer, 2019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Maharana, A., Yadav, P., and Bansal, M. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*, 2023.

Mickisch, D., Assion, F., Greßner, F., Günther, W., and Motta, M. Understanding the decision boundary of deep neural networks: An empirical study. *arXiv preprint arXiv:2002.01810*, 2020.

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

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34, 2021.

Pleiss, G., Zhang, T., Elenberg, E., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

Pooladzandi, O., Davini, D., and Mirzasoleiman, B. Adaptive second order coresets for data-efficient machine learning. In *International Conference on Machine Learning*, pp. 17848–17869. PMLR, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raju, R. S., Daruwalla, K., and Lipasti, M. Accelerating deep learning with dynamic data pruning. *arXiv preprint arXiv:2111.12621*, 2021.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. *International Conference on Learning Representations*, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Tan, H., Wu, S., Du, F., Chen, Y., Wang, Z., Wang, F., and QI, X. Data pruning via moving-one-sample-out. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Tukan, M., Zhou, S., Maalouf, A., Rus, D., Braverman, V., and Feldman, D. Provable data subset selection for efficient neural network training. *arXiv preprint arXiv:2303.05151*, 2023.

Wolf, G. W. Facility location: concepts, models, algorithms and case studies. 2011.

Xia, X., Liu, J., Yu, J., Shen, X., Han, B., and Liu, T. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Yang, S., Xie, Z., Peng, H., Xu, M., Sun, M., and Li, P. Dataset pruning: Reducing training data by examining generalization influence. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., and Kankanhalli, M. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAX0l6Cz8ub.

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S. Differentiable augmentation for data-efficient gan training. *Advances in neural information processing systems*, 33: 7559–7570, 2020.

Zheng, H., Liu, R., Lai, F., and Prakash, A. Coverage-centric coreset selection for high pruning rates. In *The Eleventh International Conference on Learning Representations*, 2022.

A. Related Work

Coreset selection methods are designed to reduce the size of training datasets while maintaining or even improving model performance (Guo et al., 2022). Popular methods can be roughly categorized into three types: geometry-based, error-based, and optimization-based methods.

Geometry-based Methods observe data distribution in the embedding space and select samples based on the distance to other samples, facilitating per-sample selection. Notably, clustering-based methods are intuitive (Feldman et al., 2020; Cohen-Addad et al., 2021; Kaddour, 2023), where Sener & Savarese (2017) utilize a greedy k-center approach, and Sorscher et al. (2022) employ the k-means algorithm to search the coreset. However, these methods are limited by their need to compute a distance matrix between samples, which leads to high time complexity. In addition, precise geometry estimation can be challenging for high-dimensional real data, e.g., natural images.

Error-based Methods primarily sorts samples by importance using various scoring mechanisms on model predictions, thereby defining the coreset's scope (Maharana et al., 2023; Huang et al., 2023). For instance, Pleiss et al. (2020) considers the Area Under the Margin (AUM) as an index to identify erroneously labeled data, while the EL2N score (Paul et al., 2021) estimates data difficulty using the L2 norm of the error vector. Coleman et al. (2019) measures the coreset from an entropy perspective. In addition to these, gradient diversity (Aljundi et al., 2019), forgetfulness (Toneva et al., 2019), and prediction uncertainty (He et al., 2023) are also taken into account. These methods are intuitive but heuristic. They lack a theoretical guarantee of the model's generalizability.

Optimization-based Methods strive to endow the coreset with superior theoretical attributes by optimizing selected data, ensuring they reflect similar training properties to the complete dataset (Tukan et al., 2023; Abbas et al., 2023). For example, (Mirzasoleiman et al., 2020) (Khan et al., 2023) and (Tan et al., 2023) emphasize gradient direction consistency after coreset selection. Killamsetty et al. (2021b) consider validation error as a metric, and others (Borsos et al., 2020; Pooladzandi et al., 2022; Yang et al., 2022) utilize influence functions (Koh & Liang, 2017) to guide optimization. Despite theoretical guarantees, optimization-based methods are computationally complex, and implementing them on large-scale dataset applications remains challenging (Raju et al., 2021).

Dicision Boundary-based Methods. Recent works in dataset pruning highlight that samples near the decision boundary contribute more to model performance (Sorscher et al., 2022; Liu et al., 2019; Li et al., 2023). However, due to the high complexity of deep neural networks' decision boundaries, it is non-trivial to directly locate those samples close to the boundary. Chen et al. (2021) emphasizes the importance of decision boundaries but ultimately chooses to use the distance to the clustering center as an estimation of the distance to the decision boundary. Ducoffe & Precioso (2018) estimate the distance to the decision boundary using adversarial samples with low classification confidence. This approach, though innovative, lacks comprehensive theoretical analysis and oversimplifies the relationship between low classification confidence and proximity to the decision boundary. Similarly, Liu et al. (2019) use model-predicted uncertainty as a proxy. Yet, prediction uncertainty is widely acknowledged as misleading (Abdar et al., 2021) and will inevitably lead to performance degradation of networks trained with the coreset. Furthermore, there is a lack of analysis on how well the selected samples reconstruct the decision boundary and the correlation between decision boundary reconstruction error and model generalization error. This paper pioneers a coreset selection method aiming directly at decision boundary reconstruction, establishing a theoretical link between decision boundary reconstruction error and model generalization ability, contributing innovatively to this research field.

B. Illustration of Samples Around the Decision Boundary

As shown in Figure 7, we showcase the boundary samples in CIFAR-10. This part of the sample tends to reflect the characteristics of the target category from an atypical perspective and contains critical information for defining decision boundaries. Thus, this subset of the sample has a key role in the model's understanding of the full picture of the target category, which directly affects the reconstruction of the decision boundary. For example, the sample for the "automobile" category contains images observed from different viewpoints rather than simply frontal views. Specifically, automobiles 0, 2, 3, and 6 are atypical observation perspectives that show more detailed features of the automobile. Even automobiles 7 and 9, which contain an automobile with common observation perspectives, have very unusual backgrounds, i.e. grassland and forest. All these particular semantic features make them samples supporting decision boundaries.

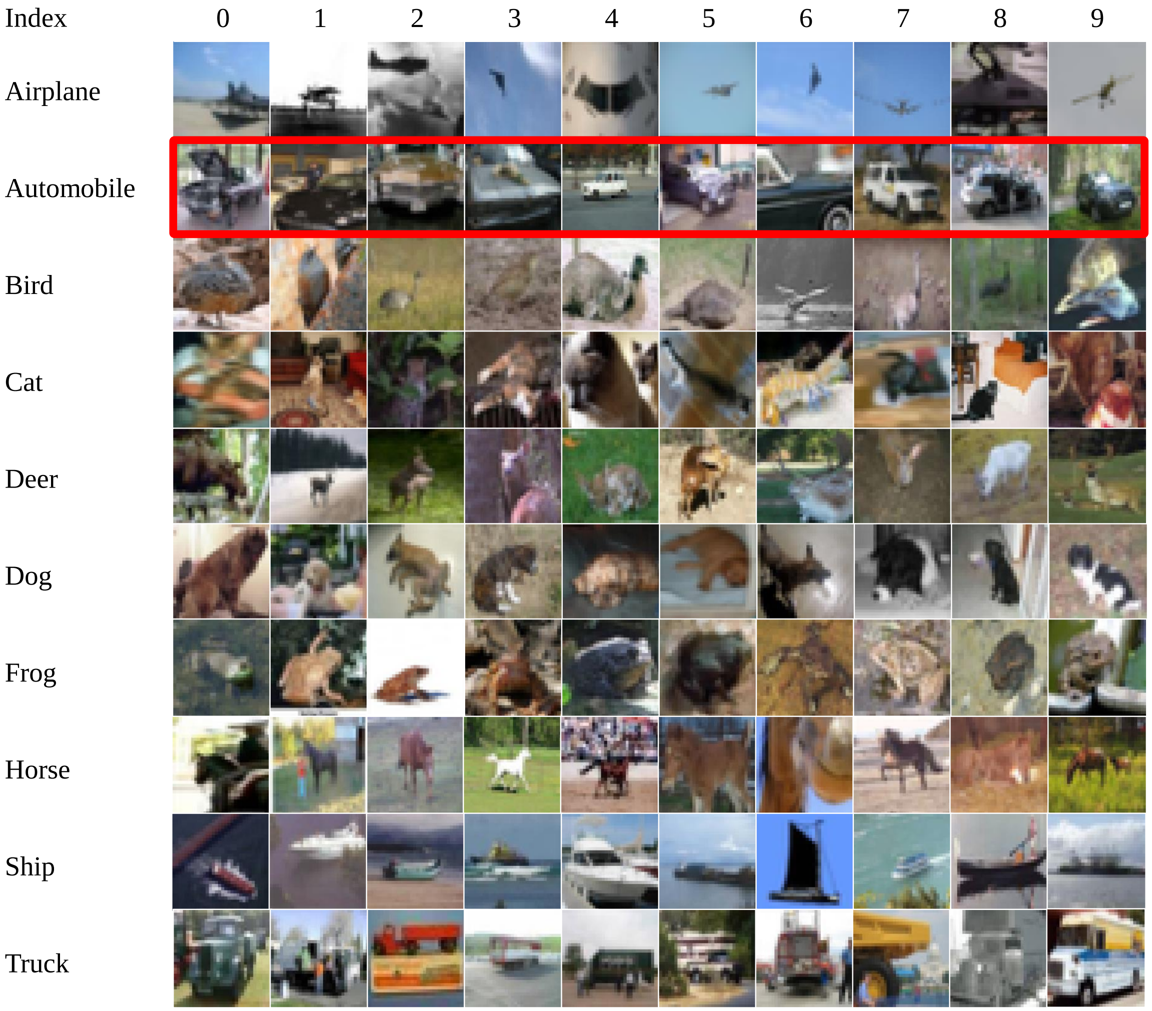


Figure 7. The top-10 samples in each category that are closest to the decision boundary.