Sparse Adversarial Perturbation-Driven Scalable Coreset Optimization

Tushar Shinde, Manasa Madabhushi
MIDAS (Multimedia Intelligence, Data Analysis and compreSsion) Lab
Indian Institute of Technology Madras, Zanzibar, Tanzania
shinde@iitmz.ac.in

Abstract

Efficient training of deep neural networks under limited data requires selecting informative subsets, or coresets, that preserve performance. Existing methods, such as DeepCore, rely on heuristics like uncertainty or gradient diversity, often overlooking adversarial vulnerabilities, which can degrade robustness under distribution shifts, corruptions, or manipulations. We introduce a unified Adversarial Sensitivity Scoring framework with three ranking strategies: Inverse Sensitivity and Entropy Fusion (ISEF), Fast Gradient Sign Method with Composite Scoring (FGSM-CS), and Perturbation Sensitivity Scoring (PSS), which leverage sparse adversarial perturbations to prioritize samples near decision boundaries. By applying single-step Sparse FGSM attacks, our methods reveal sample sensitivities with minimal computational cost. On CIFAR-10 with ResNet-18, our approaches outperform DeepFool by up to 15% in extremely sparse (1%) and low-data (10%) regimes, while matching top DeepCore methods in moderate-data settings. Bottom-ranked variants excel in sparse regimes by retaining perturbation-resilient samples, whereas top-ranked variants dominate beyond 20–30% data. These gains, achieved via efficient single-step gradients, position our framework as a scalable, deployable bridge between coreset selection and adversarial robustness, advancing data-efficient learning.

1 Introduction and Related Work

Efficient and robust learning in deep networks remains a key challenge, particularly under limited data or adversarial conditions. Coreset selection, which identifies small, informative subsets of training data, has been widely used to reduce computational costs while maintaining performance Guo et al. [2022], Shinde [2025], Gal et al. [2017], Shinde and Sharma. Traditional methods emphasize sample diversity but often neglect vulnerability to distribution shifts or adversarial manipulations Hardt et al. [2016], Hassan and Shinde.

Adversarial machine learning has shown that small perturbations can mislead deep models, emphasizing sensitivity near decision boundaries Su et al. [2019]. Leveraging this insight, we propose using sparse adversarial perturbations to guide coreset selection, prioritizing samples critical for learning efficiency and robustness. Prior work combining coreset selection and robustness either improves efficiency without robustness or vice versa Mirzasoleiman et al. [2020]. We bridge this gap with an Adversarial Sensitivity Scoring framework that uses sparse perturbations to select top-ranked samples near decision boundaries, improving both data efficiency and robustness.

In this work, we introduce a novel coreset selection strategy that leverages *sparse adversarial perturbations* to identify and prioritize training samples near the decision boundary, which are critical for both informativeness and robustness. Specifically, we employ the *Sparse Fast Gradient Sign Method (Sparse FGSM)* Goodfellow et al. [2014], a variant of FGSM that perturbs only a small subset

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Optimization for Machine Learning.

of input features, producing minimal yet highly informative perturbations. This highlights samples vulnerable to model changes and adversarial attacks, making them ideal for robust, data-efficient training under corrupted or manipulated data Hendrycks and Dietterich [2019]. The approach builds on the notion of *adversarial sensitivity*, which demonstrates that deep networks are most sensitive to small perturbations near decision boundaries Madry et al. [2017], enabling the identification of samples critical for improving robustness Wong et al. [2020].

In this direction, we propose three ranking schemes under our Adversarial Sensitivity Scoring framework, systematically categorized by their sensitivity metrics, based on the behavior of samples under Sparse FGSM attacks. We evaluate the performance of these schemes by selecting both the *top-ranked* samples (those most vulnerable or near the decision boundary) and the *bottom-ranked* samples (those least vulnerable, far from the boundary). We conduct extensive experiments on CIFAR-10, varying the selection ratio from 0.1% to 90% of the data to assess the effectiveness of our approach across different data budgets. Our results demonstrate the advantage of focusing on the *top* samples, which result in substantial improvements in classification accuracy and adversarial robustness, especially in low-data regimes.

We benchmark our method against existing coreset selection methods implemented in the DeepCore library Guo et al. [2022], including DeepFool-based selection Ducoffe and Precioso [2018]. In a range of experimental settings, our Sparse Adversarial Ranking consistently outperforms the DeepFool-based method and achieves performance comparable to the overall DeepCore library. In particular, our approach demonstrates superior adversarial robustness while maintaining high computational efficiency. This paper presents a novel framework for coreset selection that utilizes sparse adversarial perturbations to identify vulnerable and informative samples, offering a scalable solution, with potential applications in safety-critical domains.

2 Method

We construct an informative coreset guided by adversarial sensitivity using the Sparse Fast Gradient Sign Method (Sparse FGSM) to perturb training images at varying sparsity levels k. From each class, samples are ranked via three complementary techniques within our Adversarial Sensitivity Scoring framework, entropy-based, composite vulnerability, and perturbation-shift, quantifying each sample's adversarial vulnerability and contribution to model robustness. This prioritizes inputs that reveal weaknesses under distribution shifts or strategic manipulations. A Deep Neural Network is trained on CIFAR-10, and robustness is evaluated using Sparse FGSM, which perturbs only the top-k pixels by gradient magnitude Dinh et al. [2020]. Perturbed images are fed to the model to detect classification changes, capturing sensitivity to sparse adversarial noise. For each perturbation level k, class-wise probability distributions are stored for subsequent analysis.

To study data efficiency, we propose a ranking framework leveraging these distributions to guide coreset selection, extending prior methods Sener and Savarese [2017] with adversarial cues. Unlike conventional approaches, our framework emphasizes samples near decision boundaries to improve generalization. Selected coresets are used to retrain the model, and performance is compared to full-dataset baselines, evaluating both adversarial resilience and the effectiveness of informed sample selection under unreliable data conditions Shinde and Madabhushi.

2.1 Sparse FGSM Technique

To assess robustness against localized perturbations that mimic real-world data corruptions Modas et al. [2019], we adopt **Sparse FGSM**, a variant of the Fast Gradient Sign Method (FGSM) Goodfellow et al. [2014] aligned with efficient attack strategies Tramèr et al. [2017]. Unlike standard FGSM, which perturbs all input pixels, Sparse FGSM modifies only the top-k pixels with the largest gradient magnitudes, producing more realistic and less perceptible adversarial examples Hendrycks et al. [2021]. Formally, for input $x \in \mathbb{R}^n$ with true label y and loss $\mathcal{L}(f(x), y)$, standard FGSM generates: $x^{\text{adv}} = x + \epsilon \cdot \text{sign}\big(\nabla_x \mathcal{L}(f(x), y)\big)$, where ϵ is the perturbation budget and $f(\cdot)$ denotes the model's predictive function. In Sparse FGSM, perturbations are restricted to a subset $\mathcal{K} \subset \{1, \ldots, n\}$ corresponding to the k largest gradient magnitudes:

$$x_i^{\text{adv}} = \begin{cases} x_i + \epsilon \cdot \text{sign}(\nabla_{x_i} \mathcal{L}(f(x), y)), & i \in \mathcal{K}, \\ x_i, & \text{otherwise.} \end{cases} \tag{1}$$

This sparsity constraint preserves most of the input while targeting the most influential dimensions, revealing samples' vulnerability to minimal changes. By prioritizing such sensitive samples in coreset selection, we strengthen decision boundaries against strategic manipulations Hardt et al. [2016]. In our experiments, we evaluate $k \in \{1, 4, 16, 64\}$, spanning extremely sparse to moderately sparse perturbations, enabling a fine-grained analysis of model degradation and class-specific robustness patterns.

2.2 Coreset Selection Techniques

Training modern deep neural networks on large datasets is computationally intensive and often includes redundant or less informative samples. Coreset selection aims to identify a representative subset that preserves performance while reducing training cost Feldman [2019]. Standard coreset methods, however, often ignore variations in sample robustness and predictive uncertainty, which are critical for adversarial resilience and generalization.

To address this, we introduce coreset selection strategies under our *Adversarial Sensitivity Scoring* framework. Samples are systematically ranked using three complementary metrics: entropy-based (ISEF), composite vulnerability (FGSM-CS), and perturbation-shift (PSS). These metrics capture sensitivity to adversarial perturbations, predictive uncertainty, and output variability, prioritizing samples that strengthen decision boundaries against distribution shifts and manipulations Madry et al. [2017]. Concretely, the three ranking techniques are: (i) **Inverse Sensitivity and Entropy Fusion (ISEF)**, which balances perturbation sensitivity with predictive uncertainty; (ii) **FGSM-Based Composite Scoring (FGSM-CS)**, which combines adversarial vulnerability, uncertainty, and binary robustness into a weighted score; and (iii) **Perturbation Sensitivity Score (PSS)**, which quantifies the cumulative effect of localized pixel perturbations on model outputs. These metrics collectively identify samples near adversarial frontiers, enabling more efficient and robust coreset construction.

Inverse Sensitivity and Entropy Fusion (ISEF). We rank samples using a composite score that integrates adversarial sensitivity and predictive uncertainty, prioritizing points near decision boundaries to improve robustness under distribution shifts. Each sample $\mathbf{x}$ is assigned a score:

$$S(\mathbf{x}) = \alpha \cdot \tilde{S}_{\text{inv}_k}(\mathbf{x}) + (1 - \alpha) \cdot \tilde{S}_{\text{ent}}(\mathbf{x}), \tag{2}$$

where $\alpha \in [0, 1]$ balances the two components (set to $\alpha = 0.5$), and $\tilde{S}_{\text{inv}_k}$ and $\tilde{S}_{\text{ent}}$ are min-max normalized. The inverse sensitivity score S_{inv_k} assigns 1 for misclassification on clean input, $\frac{1}{k_{\min}}$ if misclassified at perturbation level $k_{\min} \in \{1, 4, 16, 64\}$, and 0 if always correctly classified. This rewards samples that are inherently fragile under minimal perturbations. The entropy score S_{ent} is the Shannon entropy of the softmax output on the clean input: $S_{\text{ent}} = -\sum_{i=1}^{C} p_i \log_{10}(p_i)$, where p_i is the predicted probability for class i and C is the number of classes, capturing uncertainty complementary to sensitivity for effective coreset selection.

FGSM-Based Composite Scoring (FGSM-CS). To quantify sample vulnerability under sparse adversarial perturbations, we define a composite score integrating three complementary components: adversarial sensitivity (S_{adv}), predictive uncertainty (S_{unc}), and binary robustness (S_{rob}):

$$S = w_{\text{adv}} \cdot S_{\text{adv}} + w_{\text{unc}} \cdot S_{\text{unc}} + w_{\text{rob}} \cdot S_{\text{rob}}, \tag{3}$$

with empirically chosen weights $w_{\text{adv}} = 0.5$, $w_{\text{unc}} = 0.3$, and $w_{\text{rob}} = 0.2$, emphasizing sensitivity while maintaining balance across factors. The adversarial sensitivity S_{adv} measures the change in softmax outputs under Sparse FGSM perturbations with $k \in \{1, 4, 16, 64\}$ pixels:

$$S_{\text{adv}} = \sum_{k} \alpha_k \cdot \|p^{(0)} - p^{(k)}\|_1, \tag{4}$$

where $p^{(0)}$ and $p^{(k)}$ are the softmax outputs for the clean and perturbed inputs, and $\alpha_k = 1/k$ emphasizes sensitivity to smaller perturbations. Predictive uncertainty S_{unc} captures the maximal softmax entropy across clean and perturbed inputs:

$$S_{\text{unc}} = \max_{k \in \{0,1,4,16,64\}} \left(-\sum_{i=1}^{C} p_i^{(k)} \log p_i^{(k)} \right), \tag{5}$$

where C is the number of classes. Binary robustness S_{rob} indicates misclassification under any perturbation:

$$S_{\text{rob}} = \begin{cases} 1, & \exists k \text{ s.t. } \hat{y}^{(k)} \neq y_{\text{true}}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

A higher composite score S identifies samples that are fragile under adversarial noise and uncertain predictions, effectively guiding coreset selection toward points critical for robust training.

Perturbation Sensitivity Score (PSS). To identify samples near decision boundaries, we define the Perturbation Sensitivity Score (PSS), measuring the sensitivity of a model's predictions to sparse, localized perturbations Biggio et al. [2013]. Let $\mathcal{K} = \{1, 4, 16, 64\}$ denote pixel perturbation sizes. For each image x_i, let $p_0(x_i)$ be the softmax output for the clean input, and $p_k(x_i)$ for the k-pixel perturbed input. The PSS is defined as:

$$\text{PSS}(x_i) = \sum_{k \in \mathcal{K}} \frac{1}{\log_2(k) + 1} \cdot \|p_k(x_i) - p_0(x_i)\|_2. \tag{7}$$

The $\log_2(k) + 1$ factor downweights larger perturbations, emphasizing sensitivity to minimal changes. High PSS values indicate samples that are fragile under small perturbations, likely near decision boundaries. Selecting such samples in the coreset improves model generalization and robustness under sparse adversarial or distribution-shifted scenarios.

3 Experimental Setup

The experiment was conducted on the Kaggle platform using an NVIDIA Tesla P100 GPU. We evaluate model performance through adversarial robustness under various perturbation conditions and coreset selection strategies.

Dataset. The CIFAR-10 dataset consists of 60,000 color images uniformly distributed across 10 distinct object categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each image has a size of 32×32 pixels with 3 color channels (RGB). The dataset is split into 50,000 training images and 10,000 test images, ensuring that the performance can be robustly evaluated.

Baseline Model. We begin by training a baseline ResNet-18 model on the CIFAR-10 training dataset. All experiments adopt the ResNet-18 architecture, consistent with prior coreset work Guo et al. [2022], ensuring a fair baseline comparison. While other architectures may yield higher baseline performance, our analysis focuses on generalizable insights, validated through extensive experiments across varying data fractions and perturbation levels.

Model Training. For the CIFAR-10 experiments, we trained a ResNet-18 architecture using stochastic gradient descent (SGD) with a batch size of 128, an initial learning rate of 0.1, momentum of 0.9, and a weight decay of 5×10^{-4}. The learning rate was scheduled with cosine annealing over 200 epochs. Standard data augmentation techniques were applied, including random cropping with padding = 4 and horizontal flips, along with CIFAR-10 normalization (mean = (0.4914, 0.4822, 0.4465), std = (0.2023, 0.1994, 0.2010)).

Adversarial Robustness Setup. To assess adversarial robustness, we generate Sparse FGSM perturbations at varying sparsity levels $k \in \{0, 1, 4, 16, 64\}$. Increasing values of k correspond to progressively higher sparsity, simulating more localized adversarial attacks on the images. For each sparsity level, we compute the predicted class probabilities of all training samples. These probabilities, along with the corresponding predicted labels and confidence scores, are stored in CSV files for subsequent analysis.

Model Training on the Coreset. We further evaluate performance by training models on selected subsets (coresets) of the CIFAR-10 dataset. The subset sizes vary across fractions of the full dataset: 0.1%, 0.5%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, and 90%. Based on the ranking scores stored in the CSV files, two types of coresets are constructed for each fraction: (1) *Top-ranked subsets*, consisting of samples identified as highly vulnerable or uncertain, and (2) *Bottom-ranked subsets*, consisting of samples deemed robust or confident. To ensure class balance and avoid selection bias, the number of samples chosen from each class is capped proportionally.

Hyperparameter Settings. For the **ISEF** method, we set the balancing factor to $\alpha = 0.5$. For **FGSM-CS**, the weighting parameters were chosen as $w_{\text{adv}} = 0.5$, $w_{\text{unc}} = 0.3$, and $w_{\text{rob}} = 0.2$. To evaluate adversarial robustness under varying levels of perturbation, we applied Sparse FGSM with sparsity levels $k \in \{1, 4, 16, 64\}$, corresponding to progressively stronger pixel perturbations.

We use classification accuracy as the primary evaluation metric to assess model performance. The accuracy is evaluated across different perturbation levels to quantify adversarial robustness.

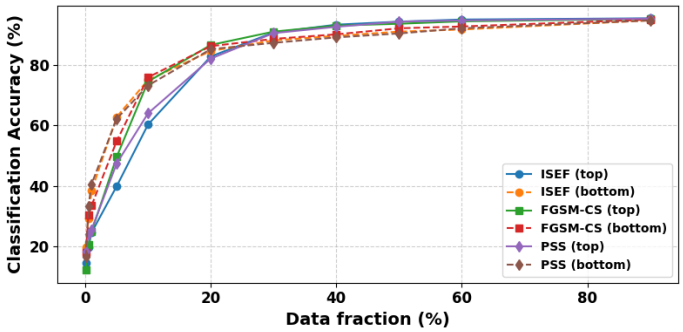

Figure 1: Performance of top vs. bottom variants across data fractions on CIFAR-10 for ISEF, FGSM-CS, and PSS, highlighting the transition where top variants overtake bottom variants around 20–30%, reflecting the shift from stable, entropy-rich samples to perturbation-sensitive samples.

4 Results and Analysis

We evaluate our Adversarial Sensitivity Scoring framework across varying data fractions, comparing top-ranked (most sensitive) and bottom-ranked (least sensitive) samples. Results demonstrate that prioritizing critical samples improves performance, particularly in low-data regimes.

Ablation Study. We evaluate the effectiveness of the proposed ranking strategies, ISEF, FGSM-CS, and PSS, on CIFAR-10 across data fractions ranging from 0.1% to 90% (Table 1). Our results demonstrate that adversarially-informed coreset selection consistently improves classification accuracy, particularly in sparse-data regimes, by mitigating label noise and highlighting samples that reveal model vulnerabilities. In extremely low-data settings (0.1%–1%), bottom-ranked samples ($\text{ISEF}_{\text{bottom}}$, $\text{PSS}_{\text{bottom}}$) dominate, indicating that stable, low-entropy points provide an effective bootstrap for initial learning. As the data fraction increases to 5%–20%, $\text{FGSM-CS}_{\text{bottom}}$ becomes most effective, suggesting that single-step gradient perturbations efficiently identify informative, non-adversarial samples that prevent overfitting. In moderate-to-high data regimes (30%–90%), top-ranked samples (ISEF_{top}) match or surpass other methods, highlighting that decision-boundary samples refine performance once a reliable model backbone is established. Figure 1 further illustrates this two-phase dynamic: bottom-ranked samples provide a robust starting point in extremely low-data scenarios, while top-ranked, perturbation-sensitive samples become critical for performance refinement as more data becomes available. This analysis underscores the importance of adaptive coreset selection tailored to data availability and model stability.

Comparison with Existing Work. We compare our adversarially-informed coreset selection strategies to traditional approaches, including DeepCore, k-Center Greedy, Random sampling, and DeepFool Ducoffe and Precioso [2018], using ResNet-18 on CIFAR-10 (Table 1). Our methods consistently outperform baseline techniques in sparse-data regimes by prioritizing hard, perturbation-sensitive samples that reveal model weaknesses, enabling efficient learning and improved robustness with reduced computational overhead. Specifically, in extremely low-data settings (0.1%–1%), $\text{PSS}_{\text{bottom}}$ and $\text{ISEF}_{\text{bottom}}$ surpass standard uncertainty- or margin-based sampling, demonstrating that targeting perturbation-sensitive points accelerates early convergence. For low-to-moderate data fractions (5%–20%), $\text{FGSM-CS}_{\text{bottom}}$ achieves the highest accuracy (75.9% at 10%), highlighting the advantage of integrating gradient and uncertainty cues to identify informative samples that prevent overfitting. In moderate-to-high data regimes (30%–90%), accuracy saturates across methods, yet top-ranked samples such as ISEF_{top} and $\text{FGSM-CS}_{\text{top}}$ maintain efficiency, reaching 95.0%–95.4% at 60%–90%, illustrating that decision-boundary samples refine performance once a stable backbone is established. Compared to DeepFool, our sparse perturbation-based selection consistently provides higher accuracy in low-data regimes (e.g., $\text{PSS}_{\text{bottom}}$: 33.3% vs. DeepFool: 22.4% at 0.5%), confirming that prioritizing samples by perturbation magnitude enables scalable, robust coreset construction suitable for large-scale or continual learning, while preserving competitive performance under abundant data.

Table 1: Classification accuracy (in %) of novel ranking techniques compared with DeepCore methods on CIFAR-10. Best, second, and third values in each column are highlighted using green shades.

Method	0.1%	0.5%	1%	5%	10%	20%	30%	40%	50%	60%	90%
Random Guo et al. [2022]	21.0	30.8	36.7	**64.5**	*75.7*	**87.1**	90.2	92.1	93.3	94.0	95.2
CD Agarwal et al. [2020]	15.8	20.5	23.6	38.1	58.8	81.3	90.8	*93.3*	94.3	94.6	*95.4*
Herding Welling [2009]	20.2	27.3	34.8	51.0	63.5	74.1	80.1	85.2	88.0	89.8	94.6
k-Center Greedy Sener and Savarese [2017]	18.5	26.8	31.1	51.4	75.8	87.0	90.9	92.8	93.9	94.1	*95.4*
Least Confidence Coleman et al. [2019]	14.2	17.2	19.8	36.2	57.6	81.9	90.3	93.1	94.5	*94.7*	95.5
Entropy Coleman et al. [2019]	14.6	17.5	21.1	35.3	57.6	81.9	89.8	93.2	*94.4*	**95.0**	*95.4*
Margin Coleman et al. [2019]	17.2	21.7	28.2	43.4	59.9	81.7	90.9	93.0	94.3	94.8	95.5
Forgetting Toneva et al. [2018]	21.4	29.8	35.2	52.1	67.0	*86.6*	**91.7**	93.5	94.1	94.6	95.3
GraNd Paul et al. [2021]	17.7	24.0	26.7	39.8	52.7	78.2	*91.2*	**93.7**	**94.6**	**95.0**	95.5
Cal Margatina et al. [2021]	**22.7**	33.1	37.8	60.0	71.8	80.9	86.0	87.5	89.4	91.6	94.7
Craig Mirzasoleiman et al. [2020]	22.5	27.0	31.7	45.2	60.2	79.6	88.4	90.8	93.3	94.2	95.5
GradMatch Killamsetty et al. [2021a]	17.4	25.6	30.8	47.2	61.5	79.9	87.4	90.4	92.9	93.2	93.7
Glister Killamsetty et al. [2021b]	19.5	27.5	32.9	50.7	66.3	84.8	90.9	93.0	94.0	94.8	**95.6**
FL Iyer et al. [2021]	*22.3*	*31.6*	38.9	60.8	74.7	85.6	91.4	93.2	93.9	94.5	95.5
DeepFool Ducoffe and Precioso [2018]	17.6	22.4	27.6	42.6	60.8	83.0	90.0	93.1	94.1	94.8	95.5
ISEF_{top}	14.5	19.8	24.6	39.9	60.3	82.8	90.7	*93.3*	94.3	**95.0**	*95.4*
$\text{ISEF}_{\text{bottom}}$	19.9	29.4	*38.4*	62.8	75.0	84.6	88.1	89.6	91.0	91.7	94.7
$\text{FGSM-CS}_{\text{top}}$	12.2	20.5	24.9	49.6	74.5	*86.6*	91.0	93.0	93.6	94.4	95.2
$\text{FGSM-CS}_{\text{bottom}}$	17.9	30.3	33.5	54.8	**75.9**	86.2	88.6	90.1	92.1	92.7	95.0
PSS_{top}	18.3	24.0	25.6	47.4	64.0	82.1	90.5	92.6	94.3	*94.7*	95.3
$\text{PSS}_{\text{bottom}}$	16.6	**33.3**	**40.6**	*62.2*	73.2	85.2	87.3	89.1	90.4	92.1	94.7

5 Conclusion and Future Work

We introduced a sparse adversarial ranking framework for data selection that efficiently identifies decision-boundary samples using single-step perturbations. Compared to iterative methods like DeepFool, our approach achieves higher accuracy with fewer samples and substantially lower computational cost. Experiments show that $\textbf{PSS}_{\textbf{bottom}}$ and $\textbf{ISEF}_{\textbf{bottom}}$ consistently outperform baselines in extremely sparse and low-data regimes, highlighting the value of prioritizing critical examples for robust, data-efficient training under distribution shifts or strategic manipulations. By focusing on informative samples rather than full datasets, our framework enables scalable training in resource-constrained settings. Its efficiency and robustness suggest broad applicability, though performance on more complex datasets or architectures warrants further evaluation.

Future Work includes integrating additional adversarial perturbation techniques and extending evaluation to larger, real-world datasets (e.g., autonomous driving, medical imaging). These directions will further assess scalability, generalization, and practical impact, reinforcing the utility of optimization-informed coreset selection in modern ML systems.

References

Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

Thu Dinh, Bao Wang, Andrea Bertozzi, Stanley Osher, and Jack Xin. Sparsity meets robustness: Channel pruning for the feynman-kac formalism principled robust deep neural nets. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 362–381. Springer, 2020.

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

Dan Feldman. Core-sets: Updated survey. In *Sampling techniques for supervised or unsupervised tasks*, pages 23–44. Springer, 2019.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer, 2022.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Aisha Hamad Hassan and Tushar Shinde. Efficient spam detection with sentence-bert using adaptive uncertainty-diversity ranking coresets. In *Women in Machine Learning Workshop@ NeurIPS 2025*.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pages 722–754. PMLR, 2021.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8110–8118, 2021b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*, 2021.

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

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9096, 2019.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

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

Tushar Shinde. An efficient and scalable framework for lightweight crop disease recognition in low-resource settings. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5534–5541, 2025.

Tushar Shinde and Manasa Madabhushi. Data-efficient and robust coreset selection via sparse adversarial perturbations. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*.

Tushar Shinde and Avinash Kumar Sharma. Scalable and efficient multi-weather classification for autonomous driving with coresets, pruning, and resolution scaling. In *ICLR 2025 Workshop on Machine Learning Multiscale Processes*.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, pages 1121–1128, 2009.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.