Improving the Efficiency of Self-Supervised Adversarial Training through Latent Clustering-based Selection

Somrita Ghosh [1] **Yuelin Xu** [1] **Xiao Zhang** [1]

Abstract

Compared to standard learning, adversarially robust learning is widely recognized to require a much larger training dataset. Recent works utilize external or synthetically generated unlabeled data in adversarial training using self-supervised learning. Despite achieving enhanced robustness, these methods typically require a considerable amount of additional data, leading to substantial memory consumption and convergence time. To address the space and computational challenges, we propose a novel Latent Clustering-based Selection scheme (LCS) to strategically select a small core subset of unlabeled data critical for obtaining better robustness. In particular, our method prioritizes selecting unlabeled data that are close to the model's decision boundary, while balancing the ratio between the boundary and the remaining data points to avoid overfitting. Our experiments show that when incorporated into self-supervised adversarial training, our LCS scheme can significantly reduce the memory and time complexities while achieving comparable model robustness.

1. Introduction

Over the past decade, it has been repeatedly confirmed that deep neural networks (DNNs) are vulnerable to adversarial perturbations (Szegedy et al., 2013). This has raised serious concerns about the reliability of DNNs in safety-critical applications and has driven numerous research into designing methods to enhance model robustness (Goodfellow et al., 2014; Papernot et al., 2016; Buckman et al., 2018; Biggio & Roli, 2018). Among them, adversarial training is regarded as one of the most effective methods to improve model robustness (Madry et al., 2017; Wang et al., 2019; Zhang et al., 2019). However, as stated by Schmidt et al. (2018), learning a robust model requires a significantly larger amount of data than that of standard learning. To address this challenge, recent studies have explored self-supervised techniques to greatly expand the training set size of adversarial training algorithms by leveraging unlabeled external (Carmon et al., 2019) or generated data (Gowal et al., 2021; Sehwag et al., 2021; Wang et al., 2023). Despite producing models with improved robust accuracy, these methods typically utilize vast amounts of additional data, suggesting the requirement of much larger hardware to store those data and a much longer training time for adversarial training to converge.

Witnessing the challenges of additional memory and computational requirements, we investigate whether the significantly large amount of utilized additional data is inevitable for achieving state-of-the-art adversarial robustness. The ultimate goal of our work is to maximize the model robustness achieved by self-supervised adversarial training algorithms by utilizing additional unlabeled data points as few as possible. Inspired by Zhang et al. (2020), which highlights the unequal importance of training examples, we argue that with limited model capacity, self-supervised adversarial learning should also focus on optimizing critical data samples near the model's decision boundary. Consequently, we propose **L**atent **C**lustering-based **S**election (LCS), a novel data selection strategy that prioritizes unlabeled data points, where the model exhibits a higher prediction uncertainty. Such a strategic data reduction scheme streamlines the self-supervised adversarial training process while attaining a comparable model's robustness against adversarial perturbations.

Contributions. We propose a Latent Clustering-based Selection (LCS) approach (Algorithm 1), aimed at reducing the volume of unlabeled data while maintaining model robustness. By strategically prioritizing boundary points, our method optimizes both efficiency and effectiveness by refining the model's decision boundary in the input regions of high uncertainty (Section 3.1). To avoid overfitting, our method strikes a balance between incorporating boundary points and leveraging the remaining points that are away from the model's decision boundary (Section 3.2). By focusing on critical unlabeled data points, our method largely reduces the computational and time complexities of self-supervised adversarial training algorithms. Our experiments on image benchmarks demonstrate that our proposed LCS scheme significantly reduces the memory consumption and

[1]CISPA Helmholtz Center for Information Security, Germany. Correspondence to: Xiao Zhang <xiao.zhang@cispa.de>.

ICML 2024 Workshop on the Next Generation of AI Safety.

the total running time of self-supervised adversarial training algorithms in diverse scenarios (Section 4), enabling the development of more scalable robust learning algorithms, particularly beneficial for resource-constrained environments.

2. Preliminaries

In this section, we introduce the background on adversarial robustness and self-supervised adversarial training.

Adversarial Robustness. Adversarial robustness captures a model's resilience to adversarial perturbations. Let $\mathcal{D}$ be the underlying distribution, from which labeled data $(\boldsymbol{x}, y)$ are i.i.d. sampled. Here, $\boldsymbol{x} \in \mathbb{R}^d$ represents an input in a d-dimensional feature space and y denotes its class label. We work with the following definition of adversarial robustness.

Definition 2.1 (Adversarial robustness)**.** Let f_θ be a classifier with parameters θ. Consider ℓ_p perturbations with strength $\epsilon \geq 0$. The adversarial robustness of f_θ is given by:

$$\mathrm{AdvRob}_\epsilon(f_\theta) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\big[\exists\, \boldsymbol{x}' \in \mathcal{B}_\epsilon(\boldsymbol{x}) \text{ s.t. } f_\theta(\boldsymbol{x}') \neq y\big],$$

where $\mathcal{B}_\epsilon(\boldsymbol{x})$ is denotes an ℓ_p-norm ball at $\boldsymbol{x}$ with ϵ radius.

When $\epsilon = 0$, adversarial robustness is equivalent to standard accuracy, i.e., $\mathrm{AdvRob}_0(f_\theta) = \mathrm{Acc}(f_\theta)$. In our experiments, we measure robust and standard accuracies using a test dataset sampeled from $\mathcal{D}$ based on Definition 2.1.

Self-Supervised Adversarial Training (SSAT). The pioneering work (Carmon et al., 2019) proposed to leverage external unlabeled data to enhance the model robustness using self-supervised learning techniques. In particular, let $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ be a labeled dataset, where each example $(\boldsymbol{x}_i, y_i)$ is i.i.d. sampled from $\mathcal{D}$. Let $D_u = \{\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_{n+N}\}$ be a set of unlabeled data. Note that the unlabeled data may follow a different input distribution from the labeled data. SSAT first trains a standard classifier $f_{\hat{\theta}}$ using the labeled dataset D_l, then assigns a pseudo-label to each unlabeled data in D_u. Consequently, a training set that consists of D_l and $D_u = \{(\boldsymbol{x}_{n+1}, \hat{y}_{n+1}), (\boldsymbol{x}_{n+2}, \hat{y}_{n+2}), \ldots, (\boldsymbol{x}_{n+N}, \hat{y}_{n+N})\}$ is prepared, where $\hat{y}_{n+i} = f_{\hat{\theta}}(\boldsymbol{x}_{n+i})$ for any $i = 1, 2, \ldots, N$. Then, both D_l and D_u are incorporated into the adversarial training framework to train robust models. To be more specific, the training objective of self-supervised adversarial training can be cast as the following optimization problem:

$$\min_\theta \sum_{(\boldsymbol{x},y)\in D_l} \mathcal{L}_{\mathrm{adv}}(\theta, \boldsymbol{x}, y) + \lambda \sum_{\boldsymbol{x}\in D_u} \mathcal{L}_{\mathrm{adv}}\big(\theta, \boldsymbol{x}, f_{\hat{\theta}}(\boldsymbol{x})\big),$$

where $\mathcal{L}_{\mathrm{adv}}$ is the adversarial loss function and $\lambda > 0$ denotes the hyperparameter that controls the trade-off between the labeled and unlabeled data distribution.

3. Proposed Data Selection Scheme

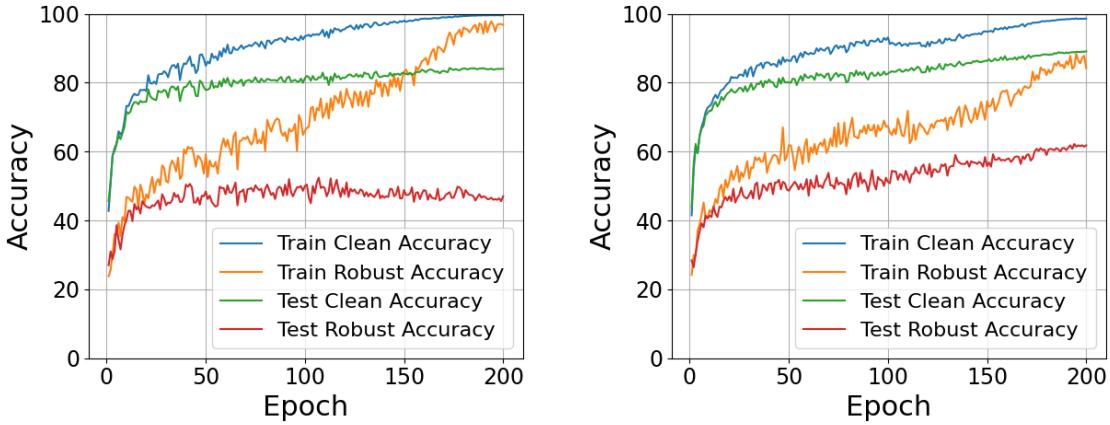
In this section, we introduce the proposed selection scheme, which is designed to improve the efficiency of SSAT by reducing the effective amount of utilized unlabeled data.

Motivation. We start by explaining the motivation of the proposed study. Schmidt et al. (2018) highlighted that adversarially robust generalization requires a significantly larger sample complexity than that required by standard generalization. Self-supervised learning, which leverages both labeled and unlabeled data, has been explored to improve adversarial robustness (Alayrac et al., 2019; Najafi et al., 2019; Zhai et al., 2019; Gowal et al., 2021; Wang et al., 2023). However, all of the aforementioned methods require a large amount of extra data to achieve a satisfactory level of robustness, suggesting inefficiencies in both memory and computation. Figure 1 illustrates the learning curves in both training and testing time for different adversarial training algorithms on CIFAR-10 with and without additional unlabeled data. For instance, the utilized unlabeled data from ImageNet are 10 times larger than the original CIFAR-10 labeled dataset, suggesting a significant increase in memory requirement to store those extra unlabeled data. In addition, when more data is utilized in adversarial training, Figure 1 shows that the best robust accuracy is achieved at a much later epoch, suggesting a much longer convergence time. We hypothesize that not all unlabeled data equally contribute to robust accuracy. Therefore, we propose to strategically select a subset of unlabeled data that contributes most to robustness enhancement if utilized in SSAT, aiming to address the memory and computational challenges while maintaining a desirable improvement on robust accuracy (see Appendix A for an illustration of the pipeline of our selection scheme).

3.1. Prioritize Boundary Unlabeled Data

To address these challenges, we aim to selectively choose a subset of this data while upholding robust accuracy. Zhang et al. (2020) stated that when faced with limited model capacity, prioritizing points more susceptible to adversarial attacks is crucial for enhancing robust accuracy. Our intuition is based on the observation that points near decision boundaries are particularly vulnerable to adversarial perturbations. Selecting unlabeled data points based on their proximity to decision boundaries can efficiently reduce the dataset size while maintaining robust accuracy to the best.

Therefore, we propose to identify the *vulnerable but valuable* data points as the unlabeled samples that are close to decision boundaries. These examples are highly susceptible to label changes caused by small input perturbations but are essential for achieving good robust generalization performance. We hypothesize that optimizing these points from the unlabeled data can significantly enhance model



(a) Vanilla Adversarial Training (b) Self-Supervised Adversarial Training

Figure 1. Comparison of learning curves on CIFAR-10 using: (a) vanilla adversarial training (Madry et al., 2017) with 50k labeled data, and (b) self-supervised adversarial training (Carmon et al., 2019) with 50k labeled data and 500k unlabeled data selected from ImageNet.

Algorithm 1 Latent Clustering based Selection (LCS)

Input: Labeled dataset D_l, unlabeled dataset D_u, number of selected data N_c, number of clusters k, parameter β
$\hat{\theta} \leftarrow$ minimize the standard classification loss on D_l
$\hat{y}_i \leftarrow$ predict pseudo label $\hat{y}_i = f_{\hat{\theta}}(\boldsymbol{x}_i)$ for each $\boldsymbol{x}_i \in D_u$
Data Selection:
  $\boldsymbol{z}_i \leftarrow$ compute embeddings $\boldsymbol{z}_i = h_{\hat{\theta}}(\boldsymbol{x}_i)$ for each $\boldsymbol{x}_i \in D_u$
  $\{\mathcal{C}_1, \ldots, \mathcal{C}_k\} \leftarrow$ apply K-means on $\{\boldsymbol{z}_i\}$ to get k clusters
  $\Delta d \leftarrow$ compute $\Delta d = |d_1 - d_2|$ for each $\boldsymbol{x}_i \in D_u$, where d_1 and d_2 are the ℓ_2 distances to the nearest two centroids
  $S_u \leftarrow$ select the top $\beta \cdot N_c$ points with the smallest Δd
  $S_u \leftarrow S_u + (1-\beta)N_c$ points randomly from $D_u \setminus S_u$
$\theta_{\text{final}} \leftarrow$ SSAT based on the adversarial loss on D_l and S_u
Output: selected dataset S_u, final model θ_{final}

robustness, achieving gains comparable to training with the entire unlabeled dataset. A straightforward approach is to select data points whose predictions from the model change the most under adversarial perturbations such as PGD attacks. However, this method is computationally intensive, negating the envisioned efficiency gains.

To identify vulnerable data points near the model's decision boundary without incurring significant computational overhead, we seek a better alternative to the aforementioned naive but inefficient method. These methods entail iterative optimization processes to pinpoint boundary points, making them less efficient, especially for large datasets. Additionally, while these methods provide valuable insights, their complexity often limits their scalability and interpretability. Therefore, we require a strategy that balances computational efficiency, scalability, and interpretability while effectively identifying boundary points. Thus, we propose Latent Clustering-based Selection (LCS) which performs K-means clustering in the latent space with an intermediate model. The pseudocode of the proposed algorithm is depicted in Algorithm 1. Initially, we group data points using k-means clustering and select those farthest from cluster centroids, assuming they are near decision boundaries. However, this assumption does not always hold, as distant points may not be near any decision boundary. To refine our approach, we select data points based on the minimal difference in distance between their two closest cluster centroids. This ensures the selected points are near decision boundaries, enhancing model robustness effectively and efficiently.

Specifically, our method first generates latent representations for unlabeled data $\{\boldsymbol{z} = h_{\hat{\theta}}(\boldsymbol{x}) : \boldsymbol{x} \in D_u\}$, where h_θ denotes the feature extractor corresponding to the mapping from the input layer to the penultimate layer of f_θ. Next, our method partitions the N unlabeled data points into k clusters $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ by minimizing the within-cluster sum of squares $\sum_{j=1}^{k} \sum_{\boldsymbol{z} \in \mathcal{C}_j} \|\boldsymbol{z} - \boldsymbol{\mu}_j\|^2$, where $\boldsymbol{\mu}_j$ is the centroid of the j-th cluster. For each data point $\boldsymbol{z}$, we calculate the Euclidean distance to each cluster's centroid $\|\boldsymbol{z} - \boldsymbol{\mu}_j\|$. Data points are selected based on the minimal difference in distance between their two closest cluster centroids $\Delta d = |d_1 - d_2|$, where d_1 and d_2 are the ℓ_2 distances to the closest and second closest centroids, respectively. As a result, the set of unlabeled inputs with the smallest Δd values, indicating the closest proximity to the model's decision boundaries, are chosen. Finally, the top N_c points from D_u with the smallest Δd values form the reduced unlabeled dataset. As will be shown in our experiments, by prioritizing such strategically selected unlabeled data points in optimizing the model's decision boundaries, self-supervised adversarial training can achieve comparable robustness with much-improved memory and computational efficiency.

Table 1. Comparison results of self-supervised adversarial training methods using varying data selection scheme and set size on SVHN (73K labeled samples). Here, **Std Acc** and **Rob Acc** stand for standard accuracy and robust accuracy evaluated on the testing dataset.

Data source	Type of Selection	Extra Set Size	Std Acc (%)	Rob Acc (%)	#Epoch	Time
External	No selection	531K	97.1	86.0	200	9.48h
	Random selection (10%)	53K	96.1	81.9	75	3.95h
	LCS (1%)	5.3K	95.2	82.7	75	3.95h
	LCS (10%)	53K	96.1	85.8	75	3.95h
	LCS (20%)	106K	96.8	82.8	75	3.95h
Generated	No selection	1M	97.4	86.3	400	18.96h
	Random selection (10%)	100K	96.3	84.5	75	3.95h
	LCS (10%)	100K	96.6	87.0	75	3.95h

3.2. Address Overfitting by Rebalancing

Zhang et al. (2020) emphasized the significance of data distribution on model performance, noting that an abundance of data far from the decision boundary can overshadow crucial boundary points. Their approach prioritizes selecting points near the decision boundary. However, focusing solely on these points can lead to overfitting due to excessive boundary points. Our experiments indicate that while including points near the boundary enhances robustness, an excessive number can cause imbalance, reducing robust accuracy. To select a balanced dataset, we refine our selection process: a proportion of the data is selected based on proximity to the boundary, while the remaining data is randomly chosen from the unlabeled set. The value of β determines the proportion of data selected from points near the boundary versus those farther away. This value depends on the amount of unlabeled data selected. If the amount of unlabeled data is small the value of β is closer to 1. Otherwise, the value of β is set to be lower to avoid overfitting to uncertain data distributions. Figure 3 in Appendix C shows how the robust accuracy varies with the value of β. For most of our experiments when selecting 10% or less of the unlabeled data, we set β to 0.6. When selecting 20% of the unlabeled data, β is set to 0.4. Line 9 of Algorithm 1 describes this balancing step, emphasizing the boundary points while reducing overfitting by incorporating randomly selected data from the remaining unlabeled distribution.

3.3. Computational Benefits

The efficiency improvement of our scheme in terms of computational aspects comes from the decreased number of total training epochs. Zhang et al. (2017) and Belkin et al. (2019) demonstrated that deep learning models typically do not overfit in standard scenarios, as both testing and training losses decrease simultaneously. This encourages training models for extended periods. However, Rice et al. (2020) identified the robust overfitting phenomenon in adversarial training, where robust accuracy on the test set deteriorates after some time despite continued improvement on the training set. Recently Wang et al. (2023) observed improved robust accuracy from extended training duration on a vast amount of generated data. This impact of different dataset sizes and the number of epochs on robust accuracy is also studied in Wang et al. (2023). Our experiments also confirm that smaller datasets require fewer training iterations to learn the underlying patterns effectively (see Table 1 and Table 2 for detailed comparisons of the effect of the training set size on the convergence rate and total running time of SSAT). With our data selection scheme significantly reducing the training set size, SSAT is thus expected to achieve optimal robust accuracy in fewer training epochs. Overshooting this optimal point may result in robust overfitting. We leverage this insight to decrease the computational complexity of SSAT by implementing early stopping.

4. Experiments

In this section, we evaluate the performance of our LCS scheme (Algorithm 1) on SVHN (Netzer et al., 2011) and CIFAR-10 (Alex, 2009). Our empirical results demonstrate that selecting a subset of unlabeled data based on their proximity to the decision boundary can yield results comparable to training with the entire dataset. Initially, we conduct experiments using unlabeled data, following the methodology outlined in Carmon et al. (2019). We selected varying ratios of the unlabeled data from $\{1\%, 10\%, 20\%\}$ and analyzed their impact on robust accuracy. To further validate the generalizability of our method, we also conduct experiments on self-supervised adversarial training methods that leverage generated data (Gowal et al., 2021), assessing whether analogous outcomes could be obtained (see Appendix B for more details of our experimental settings).

4.1. Results on SVHN

We train the intermediate model $f_{\hat{\theta}}$ using 73K labeled images from the SVHN dataset. Subsequently, we utilize the

Table 2. Comparison results of self-supervised adversarial training methods with varying data selection scheme and set size on CIFAR-10 (50K labeled samples). Here, **Std Acc** and **Rob Acc** stand for standard accuracy and robust accuracy evaluated on the testing dataset.

Data source	Type of Selection	Extra Set Size	Std Acc (%)	Rob Acc (%)	#Epoch	Time
External	No selection	500K	89.7	62.5	200	24.96h
	Random selection (20%)	100K	88.6	57.8	100	13.48h
	LCS (1%)	5K	85.6	56.4	100	13.48h
	LCS (10%)	50K	87.1	58.0	100	13.48h
	LCS (20%)	100K	88.7	60.6	100	13.48h
Generated	No selection	1M	85.7	59.4	400	49.93h
	Random selection (20%)	200K	85.4	57.2	100	13.48h
	LCS (10%)	100K	86.1	57.9	100	13.48h
	LCS (20%)	200K	85.5	58.8	100	13.48h

extra data from the SVHN dataset for data selection. Given that the additional data originates from the same distribution, there might not be a necessity to differentiate between the two when creating each batch. However, our findings indicate that an extra data to original ratio of 0.1 yields optimal results with quicker convergence. Table 1 presents the results using various proportions of the SVHN extra data with its pseudo labels. Notably, the best result is achieved when 10% of the data is selected, yielding an improvement in robust accuracy comparable to using the entire extra dataset. For comparison, we also evaluate a baseline scenario where 10% of the data is randomly selected, and the subsequent gain in robust accuracy is assessed. However, increasing the proportion of data beyond this point leads to a decline in robust accuracy, likely due to the overwhelming number of points near the decision boundary, which causes robust overfitting and decreases robust accuracy. Table 3 in Appendix B displays the results when true labels are used. Consistent with findings reported (Carmon et al., 2019), the results indicate minimal differences, confirming that most of the gains arise from the data itself rather than the labels.

Moreover, we conduct experiments using 1 million generated data points from the Denoising Diffusion Probabilistic Model (DDPM) (Gowal et al., 2021). In this case, we use an original-to-generated ratio of 0.3 for the original experiment using all 1M data but for experiments using our smaller datasets, we see that an original-to-generated ratio of 0.3 works better. Due to time and computational constraints, we limit our experiments to a single scenario of selecting 10% of the data points. Our findings, also shown in Table 1, indicate that using only 10% of the data achieves results comparable to those obtained with 1M generated data.

4.2. Results on CIFAR-10

We train $f_{\hat{\theta}}$ using the 50K labeled CIFAR-10 training images. Aligned with prior work (Carmon et al., 2019), we employ a dataset of 500K unlabeled images from TinyImageNet for data selection. For each training batch, we maintain an equal division of labeled and unlabeled data. Table 2 presents the results using various proportions of the unlabeled data with its pseudo labels. We observed that the optimal result was achieved when 20% of the unlabeled data was selected. For a better illustration of the effectiveness of our selection scheme, we also evaluated a baseline method by randomly selecting 20% of the data. Similarly to SVHN experiments, we conduct experiments using the 1M generated images from the Denoising Diffusion Probabilistic Model (DDPM) (Gowal et al., 2021). In the case of training the model with the entire 1 million generated data, we used an original-to-generated ratio of 0.3. However, when selecting a subset of the data using our selection algorithm, the best results were achieved with an original-to-generated ratio of 0.7. Our findings, as shown in Table 2, indicate that using only 20% of the data achieves results comparable to those obtained with the entire 1 million generated data. Our results confirm the effectiveness of our latent clustering-based selection scheme in improving the efficiency of SSAT algorithms while maintaining the robust accuracy of the final returned model θ_{final}.

5. Related Work

In this section, we discuss the most relevant works to ours.

Adversarial Training. Adversarial training methods have evolved in recent years, with key techniques addressing the vulnerability of machine learning models to adversarial examples. Goodfellow et al. (2014) introduced the Fast Gradient Sign Method (FGSM) which offered a computationally efficient approach by perturbing input data in the direction of the gradient. Madry et al. (2017) introduced Projected Gradient Descent (PGD) as a robust training method that employs iterative optimization to generate adversarial examples within specified constraints. The Carlini-Wagner attack by Carlini & Wagner (2017) formulated a potent adversarial

attack that considered various threat models and constraints, aiming to find minimal perturbations inducing misclassifications. Despite their effectiveness in enhancing robustness, the iterative use of these techniques, especially in the case of PGD, entails multiple forward and backward passes for each example, intensifying the computational demands during the training process. Zhang et al. (2019) argued that there exists an accuracy-robustness trade-off and then proposed TRADES to strike a balance between adversarial robustness and standard accuracy. Rice et al. (2020) claimed that overfitting to the training set significantly degrades test-time robustness across diverse datasets and perturbation models. Notably, the performance gains achieved by recent advancements in adversarial training can be effectively matched through the straightforward application of early stopping.

Robust-Self Training. The Gaussian model introduced by Schmidt et al. (2018) drew attention to a notable disparity in the requisite number of samples for robust learning compared to standard learning. They illustrated that achieving robust learning often demands a substantially larger sample complexity than conventional learning methods. This disparity in data requirements poses a challenge, especially when labeled data is limited or expensive to obtain. To address this challenge, Carmon et al. (2019) first introduced the concept of robust self-training, also known as self-supervised adversarial training, as a potential solution. This approach involves training an intermediate model using available labeled data and then utilizing this model to predict labels for unlabeled data. The newly labeled data, along with the initially labeled data, are then used to train the final model. Since then, numerous studies proposed various self-supervised methods to improve adversarial robustness by leveraging unlabeled external or synthetically-generated data (Alayrac et al., 2019; Zhai et al., 2019; Gowal et al., 2021; Sehwag et al., 2021; Wang et al., 2023). While these methods achieved the state-of-the-art robustness (Croce et al., 2020), they necessitate the use of a considerable amount of unlabeled data, which is inefficient in both memory consumption and computation.

Importance of Boundary Points. Zhang et al. (2020) argued that treating all data points equally during training is unwise, primarily due to insufficient model capacity. They proposed that data points located far from decision boundaries are more robust and secure, and should therefore be given less weight than those closer to decision boundaries. They contended that if all data points are equally emphasized during training, the model may become overwhelmed by the abundance of adversarial variations in secure data, leading to undesired robust overfitting. This scenario involves the model excessively specializing in secure data at the expense of performance on unseen or more vulnerable data. Hua et al. (2021) advocated for the application of PGD training exclusively to examples near the decision boundary, emphasizing that the primary source of robustness gains stems from these specific instances. This targeted approach aims to reduce computational complexity without compromising the robustness achieved through standard adversarial training methods. Nevertheless, obtaining high robust accuracy with restricted data remains a difficult task in machine learning. It involves striking a balance among data point significance, optimizing the use of labeled and unlabeled data, and tackling constraints imposed by model capacity.

6. Conclusion and Future Work

We illustrate the importance of data selection to improve the efficiency of SSAT while keeping robust accuracy. Emphasizing boundary data points during selection significantly enhances the model's overall robustness. Besides, we observe notable variations in results based on the source of unlabeled data, whether from the same distribution or a different one. For SVHN, we find substantial improvement in robust accuracy by adding true labels to a small set of data. This suggests manually annotating a few unlabeled data points close to the boundary can significantly enhance model robustness. Our work opens new avenues for further research, including exploring advanced data selection schemes, making the selection process more interpretable, and investigating the trade-offs between model performance, data selection, and computational costs.

References

Alayrac, J.-B., Uesato, J., Huang, P.-S., Fawzi, A., Stanforth, R., and Kohli, P. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

Alex, K. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Biggio, B. and Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156, 2018.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. In *International conference on learning representations*, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Hua, W., Zhang, Y., Guo, C., Zhang, Z., and Suh, G. E. Bullettrain: Accelerating robust neural network training via boundary example mining. *Advances in Neural Information Processing Systems*, 34:18527–18538, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Najafi, A., Maeda, S.-i., Koyama, M., and Miyato, T. Robustness to adversarial perturbations in learning from incomplete data. *Advances in Neural Information Processing Systems*, 32, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, Spain, 2011.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pp. 36246–36263. PMLR, 2023.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization (arxiv: 1611.03530). arxiv, 2017.

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. *arXiv preprint arXiv:2010.01736*, 2020.

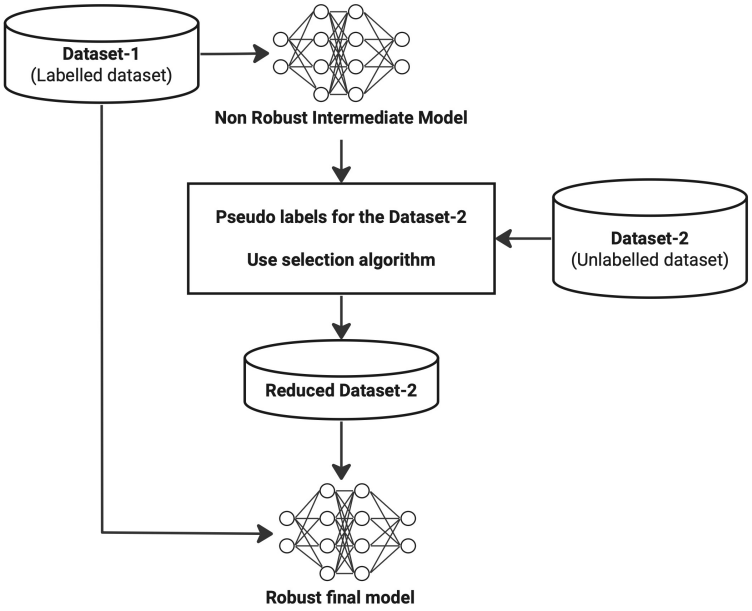


Figure 2. Overview of our approach. A model is first trained with labeled data and then used to select points from the unlabeled dataset.

A. Illustration of the Overall Pipeline

Figure 2 illustrates the overall pipeline of our data selection scheme. This is based on the robust self-training framework proposed in Carmon et al. (2019). We first train our intermediate model using the available labeled data and use this intermediate model to select a subsection of our unlabelled or generated data. The final robust training is performed on all the labeled data and a strategically selected subset of the unlabelled or generated data.

B. Experimental Details

B.1. SVHN

The SVHN dataset is naturally divided into a core training set comprising approximately 73K images and an extra training set containing around 531K images. Initially, the model is trained on the core 73K-image dataset. In our selection experiments, we investigated the effects of various percentages of boundary data, drawn from the larger 531,000-image set, on robust accuracy. Furthermore, to evaluate the generalizability of our selection algorithm, we applied it to a synthetic dataset.

Model Architecture. For all our SVHN experiments, we use Wide ResNet 16-8 (Zagoruyko & Komodakis, 2016).

Adversarial Training: We generate adversarial examples using PGD attacks exactly as implemented in Zhang et al. (2019), with step size $\alpha = 0.007$, 10 PGD attack iterations and ℓ_∞ perturbation magnitude $\epsilon = 0.015$.

Optimizer Configuration. Hyperparameters are set the same as in Carmon et al. (2019) except for the number of epochs. For SVHN, we use a training batch size of 128.

Number of epochs. All of our experiments that utilize all of the extra 531K data are run for 200 epochs and the experiments with 1M generated data are run for 400 epochs. To prevent overfitting and reduce computational complexity, we run our experiments with selected data for 75 epochs with early stopping.

Attack Evaluation. For the attack evaluation to calculate the robust accuracy, we keep the parameters similar to that of Carmon et al. (2019) for better comparison. We use step size $\alpha = 0.005$, number of attack steps $K = 100$, and number of restarts $\rho = 10$. We evaluate models at $\epsilon = 0.015$, which is the same as the value we used during training.

Table 3. Comparison results of self-supervised adversarial training with varying data selection scheme on SVHN, where the additional data are assigned with ground-truth SVHN class labels (instead of pseudo labels).

Type of Selection	Extra Set Size	Std Acc (%)	Rob Acc (%)	#Epoch
No selection	531K	97.5	86.4	200
Random selection (10%)	53K	96.3	82.7	75
LCS (1%)	5.3K	95.6	82.9	75
LCS (10%)	53K	96.7	86.3	75
LCS (20%)	106K	96.4	82.8	75

B.2. CIFAR-10

The CIFAR-10 dataset has 50K labeled images. To obtain extra data, we use the 80 Million Tiny Images (80M-TI) dataset, of which CIFAR-10 is a manually labeled subset. However, most images in 80M-TI do not correspond to CIFAR-10 image categories. Carmon et al. (2019) used an 11-way classifier to distinguish CIFAR-10 classes and an 11th "non-CIFAR-10" class using a WideResNet 28-10 model. For each class, they selected an additional 50K images from 80M-TI using the model's predicted scores to create a 500K pseudo-labeled dataset which we use in our experiments. We train the intermediate model with 50K labeled data and use this model to select data from the 500K pseudo-labeled data or a synthetic dataset. We perform the same experiments as in SVHN except that we do not have the true labels of the additional unlabeled dataset.

Model Architecture. For all our CIFAR-10 experiments, we use Wide ResNet 28-10 (Zagoruyko & Komodakis, 2016).

Adversarial Training. Similar to SVHN experiments, we set step size $\alpha = 0.007$, PGD attack iterations as 10 and ℓ_∞ perturbation magnitude $\epsilon = 0.031$.

Optimizer Configuration. Hyperparameters used are same as in Carmon et al. (2019) except for the number of epochs. Here, we use a training batch size of 256.

Number of epochs. The experiments using all the 500K pseudo-labeled data are run for 200 epochs and using 1M generated data are run for 400 epochs. On the other hand for our experiments with limited data, we get the best results at 100 epochs where we early stop the training process.

Attack Evaluation. The attack evaluation is conducted similar to that of Carmon et al. (2019) for better comparisons. We use the step size $\alpha = 0.01$, number of attack steps $K = 40$, and number of restarts $\rho = 5$. We evaluate models at $\epsilon = 0.031$, which is the same as the perturbation magnitude we used during training.

C. Other Experiments

C.1. Ground-Truth Label

We conduct additional experiments to test the performance of our data selection scheme on SVHN with pseudo labels replaced by ground-truth labels for the extra data. Table 3 shows the comparison results. With ground-truth labels, our LCS scheme can achieve similar robust accuracy by selecting just 10% extra data compared with no selection result.

C.2. Effect of Hyperparameter β

We study how different ratios of points near and far from the boundary affect the robust and standard accuracy. Figure 3 presents the results of varying β on CIFAR-10 data, with the extra data taken by selecting 20% from the 1 million images generated by Denoising Diffusion Probabilistic Model (DDPM). The best performance is achieved when β is set as 0.4.

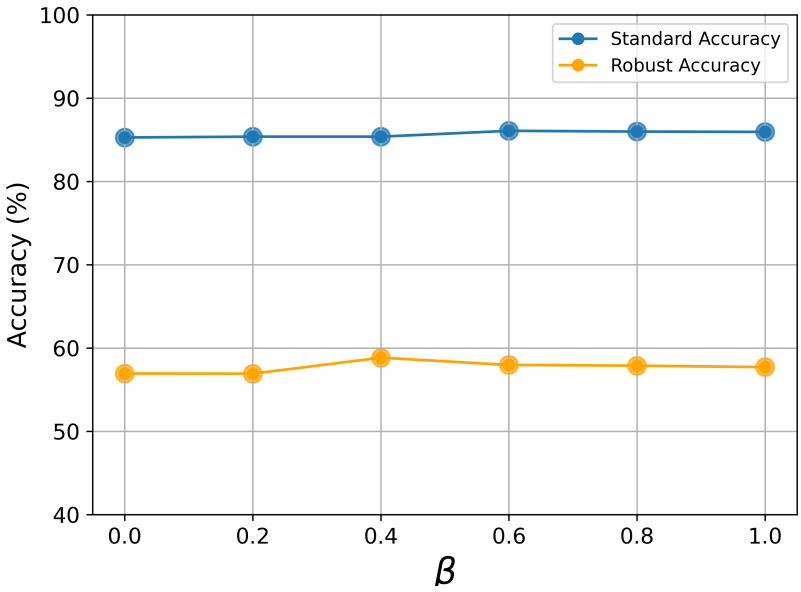


Figure 3. Variation in Standard and Robust Accuracy with respect to different β values.