SAAL: Sharpness-Aware Active Learning

Yoon-Yeong Kim [*1] **Youngjae Cho** [*2] **JoonHo Jang** [2] **Byeonghu Na** [2] **Yeongmin Kim** [2] **Kyungwoo Song** [3] **Wanmo Kang** [2] **Il-Chul Moon** [2 4]

Abstract

While deep neural networks play significant roles in many research areas, they are also prone to overfitting problems under limited data instances. To overcome overfitting, this paper introduces the first active learning method to incorporate the sharpness of loss space into the acquisition function. Specifically, our proposed method, Sharpness-Aware Active Learning (SAAL), constructs its acquisition function by selecting unlabeled instances whose perturbed loss becomes maximum. Unlike the Sharpness-Aware learning with fully-labeled datasets, we design a pseudo-labeling mechanism to anticipate the perturbed loss w.r.t. the ground-truth label, which we provide the theoretical bound for the optimization. We conduct experiments on various benchmark datasets for vision-based tasks in image classification, object detection, and domain adaptive semantic segmentation. The experimental results confirm that SAAL outperforms the baselines by selecting instances that have the potentially maximal perturbation on the loss. The code is available at https://github.com/YoonyeongKim/SAAL.

1. Introduction

A large-scale dataset is important because its wide coverage in the data space provides the generalization capability (Bartlett & Mendelson, 2002). If a deep learning model is trained with only a few data instances, the flexibility of the learning model becomes prone to overfitting (Keskar et al., 2017; Neyshabur et al., 2017). To overcome this problem of small datasets, *active learning* has been developed to iteratively select key data instances through acquisition functions, which aims at the efficient use of the limited budget for annotations from oracle (Cohn et al., 1996). This efficient usage is a difficult challenge because the value of data instance needs to be anticipated without any supervision in prior to the oracle query (Dasgupta & Hsu, 2008).

This paper proposes a new active learning algorithm, named Sharpness-Aware Active Learning (SAAL), that proposes an acquisition function by reducing the potential sharpness of loss surface after learning an instance, which is an acquisition candidate. While we are inspired by Sharpness-Aware Minimization, or SAM (Foret et al., 2020), which minimizes the maximally perturbed loss of training datasets for the flat loss surface, the adaptation of SAM to active learning requires anticipation of the sharpness without a label on a data instance. Therefore, we utilize pseudo-labels predicted by the current classifier.

This utilization of pseudo-labels calls for theoretical investigations, so we show that pseudo-labeling becomes the lower bound of the maximally perturbed loss w.r.t. ground-truth label, so such utilization can be a part of acquisition functions. Also, we theoretically derive the upper bound of the proposed acquisition score of SAAL, which includes the loss, the norm of gradients, and the first eigenvalue of loss Hessian matrix. Among the three terms of the upper bound, the loss and the gradient terms are widely used acquisition score for active learning, which captures the model change by acquiring the instance (Yoo & Kweon, 2019; Ash et al., 2020; Settles et al., 2007). Meanwhile, the first eigenvalue, which is newly considered by SAAL, is connected to the loss sharpness (Keskar et al., 2017), and this added term is related to the generalization of the model. We summarize our contributions in three aspects.

- SAAL is the first active learning framework to consider the loss sharpness in its acquisition function.
- We prove the theoretic bound of the acquisition score by utilizing the pseudo-label in SAAL.
- SAAL performs better than baseline models in various benchmarks and tasks.

[*]Equal contribution [1]Agency for Defense Development, AI Autonomy Technology Center (ADD, AIA Center) [2]KAIST, South Korea [3]Yonsei University, South Korea [4]Summary.AI. Correspondence to: Il-Chul Moon <icmoon@kaist.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Background

2.1. Notations

We assume a classifier parameterized by θ as $f_\theta\colon \mathbb{R}^d \to \mathbb{R}^{|\mathscr{Y}|}$; where d is the dimension of data instance, x; and $\mathscr{Y}$ is the set of candidate classes. There are two datasets: a dataset with labels, $\mathscr{X}_L$, and the other unlabeled dataset, $\mathscr{X}_U$. We denote the acquisition function of active learning as $f_{acq}\colon \mathbb{R}^d \to \mathbb{R}$, which receives a data instance as input, and which calculates the informativeness, or the acquisition score. The loss of a data instance, x, w.r.t. the given label y is represented as $l(x, y; \theta) \coloneqq l_{\text{CE}}(\sigma(f_\theta(x)), y)$, where l_{CE} is cross-entropy loss, and $\sigma(\cdot)$ is a softmax function. The total loss of a dataset, S, is represented as $L_S(\theta) = \frac{1}{N}\sum_{i=1}^{N} l(x_i, y_i; \theta)$, where $S = \{(x_i, y_i) | i = 1, ..., N\}$. Lastly, we define the pseudo-label, $\hat{y} = \text{argmax}_{j\in\mathscr{Y}}\, \sigma(f_\theta(x))_j$; and we denote the ground-truth label as $\bar{y}$.

2.2. Active Learning

There are several active learning scenarios that differ by the setting of data accessibility: 1) membership-query synthesis (Angluin, 1988; 2004), 2) stream-based active learning (Atlas et al., 1989; Cohn et al., 1994), and 3) pool-based active learning (Lewis & Gale, 1994). This paper focuses on pool-based active learning: an unlabeled and large dataset becomes a data pool, and the active learner sequentially selects the informative instances by acquisitions.

There are three research directions in Pool-based active learning. **1) Uncertainty-based active learning** adopts the acquisition function, f_{acq}, to calculate the uncertainty of each unlabeled instance with regard to the current deep learning model, and an oracle provides the ground-truth label of the selected unlabeled instances with the highest uncertainty. Since the acquisition score is usually calculated for an unlabeled instance, $x_u \in \mathscr{X}_U$, w.r.t. the current model, f_θ, it is expanded as $f_{acq}(x_u; f_\theta)$, resulting in the selection rule as the below.

$$\mathscr{X}_S = \underset{\mathscr{X}'_S \subset \mathscr{X}_U}{\text{argmax}} \sum_{x_u \in \mathscr{X}'_S} f_{acq}(x_u; f_\theta) \tag{1}$$

Entropy, which is denoted as $f^{Ent}_{acq}(x_u; f_\theta) = \mathbb{H}[f_\theta(x_u)] = -\sum_j \sigma(f_\theta(x_u))_j \log \sigma(f_\theta(x_u))_j$, or variation ratio, which is denoted as $f^{Var}_{acq} = 1 - \max_j \sigma(f_\theta(x_u))_j$, are the most widely used methods for calculating uncertainty (Shannon, 1948; Freeman, 1965).

Recently, additional networks are used to approximate the uncertainty of each instance. Learning Loss for Active Learning (LL4AL) (Yoo & Kweon, 2019) trains the loss prediction module, f_{LPM}, which takes the hidden feature maps as input and predicts the expected loss as output. Then, LL4AL constructs the acquisition functions $f^{LL4AL}_{acq}(x_u) = f_{LPM}(f^k_\theta(x_u)|_{k=1,...,K})$, where f^k_θ is the k-th hidden feature map. Variational Adversarial Active Learning (VAAL) (Sinha et al., 2019) trains a discriminator, f_{dis}, which takes a data instance as input and discriminates whether the instance belongs to the labeled dataset or the unlabeled dataset. Then, VAAL calculates the probability of x_u belonging to the unlabeled dataset, $\mathscr{X}_U$, as the acquisition score, i.e., $f^{VAAL}_{acq}(x_u) = f_{dis}(x_u)$.

2) Diversity-based active learning, such as Coreset approach (Sener & Savarese, 2018), selects instances that represent the whole distribution of unlabeled instances, by solving a mixed integer programming. **3) Hybrid-based active learning** is proposed to select the uncertain instances in a diverse way. In BADGE (Ash et al., 2020), the acquisition function is calculated as the gradient embedding of x_u w.r.t. the parameter of the last fully connected layer, θ_{out}, that is $f^{BADGE}_{acq}(x_u) = \frac{\partial}{\partial \theta_{out}} l(x_u, \hat{y}_u; \theta)$, where $\hat{y}_u$ is the pseudo-label of x_u. Then, this embedding becomes an input to the k-means++ seeding algorithm (Arthur & Vassilvitskii, 2006).

2.3. Sharpness-Aware Minimization (SAM)

As an independent research direction from active learning, there is an increasing investigation on the flatness (or sharpness) of loss response surfaces, and their corresponding optimization because the flat minima is confirmed to have deep connection to the generalization performance of neural networks (Jiang et al., 2019).

Sharpness-Aware Minimization (SAM) is an optimizer for training the deep neural network (Foret et al., 2020) to weigh the importance of flat minima. Denoting the loss on the dataset S w.r.t. the current parameter θ as $L_S(\theta)$, the optimization objective of SAM is minimizing the maximally perturbed loss with the regularization on the parameter, as below.

$$\min_\theta \max_{\|\epsilon\| \le \rho} L_S(\theta + \epsilon) + \gamma \|\theta\|_2^2 \tag{2}$$

Here, γ is a hyperparameter that controls the magnitude of the effect of regularization, ϵ is the perturbation to the parameter, and ρ defines the possible range of the perturbation.

This maximally perturbed loss can be decomposed as $\max_{\|\epsilon\|\le\rho} L_S(\theta+\epsilon) = (\max_{\|\epsilon\|\le\rho} L_S(\theta+\epsilon) - L_S(\theta)) + L_S(\theta)$, interpreted as the sharpness term (first term of the RHS) and the classification loss term (second term of the RHS). Hence, SAM minimizes the loss sharpness as well as the classification loss value. This optimization is a max-min problem. The inner maximization problem is solved by finding $\epsilon^* = \text{argmax}_{\|\epsilon\|\le\rho} L_S(\theta+\epsilon)$. By deriving Taylor expansion of $L_S(\theta+\epsilon)$ w.r.t. θ around 0, and by introducing a dual norm problem, the ϵ^* is approximated as follows,

with $\frac{1}{p}+\frac{1}{q}=1$.

$$\epsilon^* \approx \rho \cdot \text{sign}(\nabla_\theta L_S(\theta)) \frac{|\nabla_\theta L_S(\theta)|^{q-1}}{(\|\nabla_\theta L_S(\theta)\|_q^q)^{1/p}} \tag{3}$$

After solving the inner maximization using ϵ^*, the minimization problem is solved by obtaining the gradient, while excluding the Hessian term, as below.

$$\nabla_\theta \max_{\|\epsilon\|\leq\rho} L_S(\theta+\epsilon) \approx \nabla_\theta L_S(\theta)|_{\theta+\epsilon^*} \tag{4}$$

3. Method

3.1. Motivation

According to SAM (Foret et al., 2020), the loss of the population dataset, $\mathscr{D}$, is upper bounded by the maximally perturbed loss of the training dataset, $\mathscr{X}$. From the perspective of active learning, the training dataset is decomposed into the labeled dataset, $\mathscr{X}_L$, and the unlabeled dataset, $\mathscr{X}_U$, i.e., $\mathscr{X} = \mathscr{X}_L \cup \mathscr{X}_U$. Hence, the upper bound can be decomposed as below, with $\pi_L = \frac{|\mathscr{X}_L|}{|\mathscr{X}|}$ and $\pi_U = \frac{|\mathscr{X}_U|}{|\mathscr{X}|}$.

$$\begin{aligned} L_{\mathscr{D}}(\theta) &\leq \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}}(\theta+\epsilon) + \gamma\|\theta\|_2^2 && (5)\\ &\leq \pi_L \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_L}(\theta+\epsilon) + \pi_U \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_U}(\theta+\epsilon) \\ &\quad + \gamma\|\theta\|_2^2 && (6)\\ &=: L_{\mathscr{X}}^{SAAL} && (7) \end{aligned}$$

Since the population loss, $L_{\mathscr{D}}(\theta)$, is never accessible, we instead access the upper bound denoted in Eq. 7, which is represented as $L_{\mathscr{X}}^{SAAL}$, and train our network to minimize the upper bound. Among the three terms of $L_{\mathscr{X}}^{SAAL}$, the first term and third term, $\pi_L \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_L}(\theta+\epsilon)+\gamma\|\theta\|_2^2$, will be minimized if we use SAM optimizer. Then, the remaining second term, $\pi_U \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_U}(\theta+\epsilon)$, becomes the key component for our optimization in the sharpness-aware active learning scenario. During the acquisition iterations, we select unlabeled instances, $x_u \in \mathscr{X}_U$, with maximally perturbed losses.

As a consequence of acquiring instances with maximally perturbed losses, the acquired instances contribute to $L_{\mathscr{X}_L}$, not $L_{\mathscr{X}_U}$ anymore. Therefore, we directly reduce $L_{\mathscr{X}_U}$ (eventually, $L_{\mathscr{X}}^{SAAL}$) by removing its maximally contributing instances. Moreover, the acquired instances will be labeled, and SAM will optimize $L_{\mathscr{X}_L}$, which becomes the reduction of $L_{\mathscr{X}}^{SAAL}$, as well. These reductions of $L_{\mathscr{X}}^{SAAL}$ will reduce $L_{\mathscr{D}}(\theta)$ because of the above bound.

Comparison to Semi-Supervised Learning Our proposed active learning algorithm is not the only way to decrease the loss of the unlabeled dataset, $\mathscr{X}_U$. Traditional semi-supervised learning (SSL) is another approach that utilizes $L_{\mathscr{X}_U}(\theta)$ during model training. However, it should be noted that SSL does not guarantee to minimize the upper bound, $L_{\mathscr{X}}^{SAAL}$. SSL minimizes the average of unlabeled dataset loss instead of the maximum perturbed loss. Hence, it is hard to guarantee that SSL will contribute to minimizing the generalization error without prior knowledge on label distribution (Ben-David et al., 2008). We can categorize the SSL approach as three ways (Berthelot et al., 2019; Zhu, 2005), which are consistency regularization (Laine & Aila, 2016; Sajjadi et al., 2016), entropy minimization (Cireşan et al., 2010; Lee et al., 2013), and traditional regularization, such as weight decay (Zhang et al., 2018a;b). First, consistency regularization and entropy minimization completely depend on the pseudo-label, and an incorrect pseudo-label might increase the generalization error. Second, the worst-case or hardest instances might have incorrect pseudo-label. In other words, SSL, training the model with an incorrect pseudo-label, might fail to model the maximum perturbed loss. Third, the minimization of maximum perturbed loss is an independent approach to the previous semi-supervised learning methods, such as traditional regularization as well as consistency and entropy minimization. This aspect makes SAAL to be potentially compatible with SSL.

Algorithm 1 Sharpness-Aware Active Learning

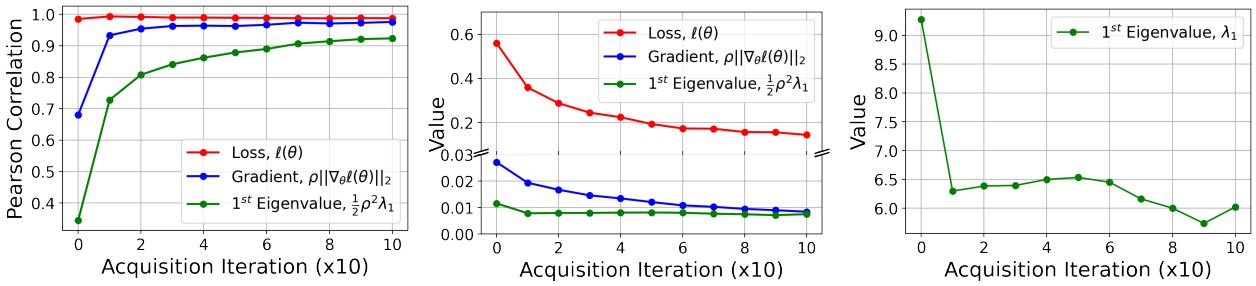
Input: Labeled dataset $\mathscr{X}_L^0$, Unlabeled dataset $\mathscr{X}_U^0$, Classifier f_θ
Initially train f_θ by the cross-entropy loss of $\mathscr{X}_L^0$
for $j = 0, 1, 2, \ldots$ **do**
  Randomly sample $\mathscr{X}_U^{pool} \subset \mathscr{X}_U^j$
  for $x_u \in \mathscr{X}_U^{pool}$ **do**
    Calculate $f_{acq}^{SAAL}(x_u; f_\theta)$ as Eq. 8
  end for
  $\mathscr{X}_S = \text{argmax}_{\mathscr{X}_S' \subset \mathscr{X}_U^{pool}} \sum_{x_u \in \mathscr{X}_S'} f_{acq}^{SAAL}(x_u; f_\theta)$
  Query the label of $\mathscr{X}_S$ to oracle
  Update the labeled dataset, $\mathscr{X}_L^{j+1} = \mathscr{X}_L^j \cup \mathscr{X}_S$
  Update the unlabeled dataset, $\mathscr{X}_U^{j+1} = \mathscr{X}_U^j \setminus \mathscr{X}_S$
  Train f_θ by the cross-entropy loss of $\mathscr{X}_L^{j+1}$
end for

3.2. Sharpness-Aware Active Learning

SAAL selects instances with high perturbed losses under some perturbation on the model parameters, θ. Hence, our acquisition function is as follows:

$$f_{acq}^{SAAL}(x_u; f_\theta) = \max_{\|\epsilon\|\leq\rho} l(x_u, \hat{y}_u; \theta+\epsilon), \tag{8}$$

where l is the cross-entropy loss function, and θ is the current model parameter. Algorithm 1 describes the overall process of SAAL. Since our acquisition function is calculated for the unlabeled instances, there comes a problem when calculating the maximally perturbed loss function, which requires labels. Hence, we use a pseudo-label, $\hat{y}_u$, for



(a) Correlation between f_{acq}^{SAAL} and upper bound terms

(b) Magnitude of the upper bound terms

(c) Detailed view of the first eigenvalue, λ_1

Figure 1. Correlation and magnitude of f_{acq}^{SAAL}'s upper bound terms; task loss, gradient norm, and 1^{st} Eigenvalue of loss Hessian matrix.

the loss calculation.

To provide the validity of utilizing pseudo-labels, we provide Theorem 3.1, which explains the relation between the maximally perturbed losses which are calculated with a pseudo-label and with a ground-truth label, respectively. The proof of Theorem 3.1 is given in Appendix A.9.1.

Theorem 3.1. *For a data instance x, let $\hat{y}$ be the pseudo-label predicted by the network f_θ and $\bar{y}$ be the ground-truth label. Then, the maximally perturbed loss calculated with $(x, \hat{y})$ is a lower bound of the maximally perturbed loss calculated with $(x, \bar{y})$; with a non-negative margin, δ_x, as the below:*

$$\max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon) \leq \max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon) + \delta_x. \quad (9)$$

Next, Proposition 3.2 shows that the inequality of Eq. 9 has zero margin under a mild condition. The proof of Proposition 3.2 is given in Appendix A.9.2.

Proposition 3.2. *For a data instance x and the corresponding pseudo-label $\hat{y}$, let $\hat{\epsilon}$ be the maximal perturbation over the parameters w.r.t. the loss $l(x, \hat{y}; \theta + \epsilon)$. If the perturbed network, $f_{\theta+\hat{\epsilon}}$, keeps the predicted label as the same as the label predicted from the original network, f_θ; then the maximally perturbed loss calculated with $(x, \hat{y})$ is a lower bound of the maximally perturbed loss calculated with $(x, \bar{y})$, as the below:*

$$\max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon) \leq \max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon). \quad (10)$$

Theorem 3.1 and Proposition 3.2 provide that the perturbed loss with the pseudo-label, $\max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon)$ becomes the lower bound of the ground-truth label, so the maximization of pseudo-label loss would indirectly increase the perturbed loss with the ground-truth label, which achieves the goal of f_{acq}^{SAAL}. On the other hand, the gap between two terms originates from the scenario of active learning, which inevitably utilizes the pseudo-label.

From (Foret et al., 2020), Eq. 3 becomes the maximal perturbation for a batch in training as the closed-form solution. However, this approach becomes inadequate for acquisition setting because the acquisition is determined by an instance, not by a batch set. Therefore, we need to calculate the closed-form optimization per instance, as below.

$$\epsilon^* \approx \rho \cdot \text{sign}(\nabla_\theta l(x_u, \hat{y}_u; \theta)) \frac{|\nabla_\theta l(x_u, \hat{y}_u; \theta)|^{q-1}}{(\|\nabla_\theta l(x_u, \hat{y}_u; \theta)\|_q^q)^{1/p}} \quad (11)$$

In the next step, we calculate the perturbed loss in direction to ϵ^*, and use it as the acquisition score:

$$f_{acq}^{SAAL}(x_u; f_\theta) = l(x_u, \hat{y}_u; \theta + \epsilon^*) \quad (12)$$

3.3. Connection to Recent Active Learning Algorithms

Here, we theoretically derive the upper bound of the acquisition score of SAAL, and this derivation shows the connection to the recent active learning algorithms as well as the generalization ability. We provide Theorem 3.3 as below. The proof of Theorem 3.3 is given in Appendix A.9.3.

Theorem 3.3. *The acquisition function, f_{acq}^{SAAL}, of Eq. 8 is upper bounded by:*

$$f_{acq}^{SAAL}(x_u; f_\theta) \leq \underbrace{l(\theta)}_{\textit{Task Loss}} + \underbrace{\rho\|\nabla_\theta l(\theta)\|_2}_{\textit{Gradient Norm}} + \underbrace{\frac{1}{2}\rho^2\lambda_1}_{1^{st}\textit{ Eigenvalue}} + \max_{\|v\| \leq 1} O(\rho^2 v^3) \quad (13)$$

Theorem 3.3 derives the upper bound of the acquisition score of SAAL, which consists of the task loss, the gradient norm, and the first eigenvalue of the loss Hessian matrix. Since we are selecting instances that have a high value of f_{acq}^{SAAL}, the selection refers that we are also selecting instances that have high values of the loss, $l(\theta)$, and the magnitude of the gradient embedding, $\|\nabla_\theta l(\theta)\|_2$, which are connected to LL4AL (Yoo & Kweon, 2019) and BADGE (Ash et al., 2020), respectively. Furthermore, SAAL considers the first eigenvalue of the loss Hessian matrix, w.r.t. the current model parameters, denoted as λ_1. The importance of

the first eigenvalue for generalization is widely studied, that is the first eigenvalue is used as the indicator of the sharpness of the loss surface (Keskar et al., 2017; Zhuang et al., 2022; Kaur et al., 2022). Hence, the selected instances by SAAL might contribute to the generalization of the model.

Figure 1a shows that there exists a positive correlation between our acquisition score, f_{acq}^{SAAL}, and the three terms of upper bound. At the same time, those three terms are not identical, which means that they are providing different information. By selecting the instances with the high acquisition score of SAAL, f_{acq}^{SAAL}, we are selecting instances that have high values of the loss, gradient norm, and the first eigenvalue. Also, Figure 1b shows the value of the three terms of upper bound. Interestingly, as the acquisition iterations proceed, not only the loss and the gradient value, but the first eigenvalue gets smaller. The change of the value of the first eigenvalue is more noticeable in Figure 1c, which plots the value of λ_1 without the scaling term of $\frac{1}{2}\rho^2$. This indicates that SAAL leads the model to a flat minima, which results in better generalization performances.

4. Results

4.1. Image Classification

Experiment Setting We conduct our experiment on Fashion-MNIST (Fashion) (Xiao et al., 2017), SVHN (Netzer et al., 2011), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009). We adopt ResNet-18 (He et al., 2016) as a backbone of our classifier. We train the network for 50 epochs after each acquisition step, using Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.001; or SAM optimizer (Foret et al., 2020) with a learning rate of 0.001 for Fashion, SVHN, CIFAR-10, and 0.1 for CIFAR-100. This comparison of optimizer choice provides the ablation between SAM and SAAL since the two share the pursuit of flatness from the loss curve. In ImageNet experiment, we follow the above settings besides 500 training epochs after each acquisition step by the Adam optimizer with 0.001 learning rate. We replicated three times for each setting. We followed the settings of the prior work (Kim et al., 2021), which assumes a very low amount of allowed budget [1]. We provide more details in Appendix A.4.

Baselines We compared the performance of SAAL with Random, Entropy (Shannon, 1948), Coreset (Sener & Savarese, 2018), Learning Loss for Active Learning (LL4AL) (Yoo & Kweon, 2019), Variational Adversarial Active Learning (VAAL) (Sinha et al., 2019), and BADGE (Ash et al., 2020). In addition, we compare our strategy with ProbCover (Yehuda et al., 2022), utilizing the features of unlabeled instances from self-supervised pretrained model. BADGE adopts k-means++ seeding algorithm to introduce diversity on the acquisition, and we also provide an experimental result with diversity following the same practice from BADGE. Specifically, after calculating our acquisition function using Eq. 8, we implement k-means++ seeding algorithm with the acquisition score as an input, and we report such variations on Table 1.

[1] Section 4.2 provides an ablation study on the budget factor.

Quantitative Analysis Table 1 indicates that SAAL outperforms the baselines in seven out of eight combinations of experiments. The advantage of SAAL becomes obvious when we use the Adam optimizer, rather than the SAM optimizer. We conjecture that this gain for Adam optimizer originates from Eq. 7, which motivates SAAL in modeling the expected flat local minima after acquisitions. Recall that our inaccessible goal, $L_{\mathscr{D}}(\theta)$, is upper bounded by $\pi_L \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_L}(\theta+\epsilon) + \pi_U \max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_U}(\theta+\epsilon)$, as we discussed in Section 3.1. When using Adam optimizer, the first term, $\max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_L}(\theta+\epsilon)$, in the upper bound is weakly optimized compared to using SAM optimizer, which we will present qualitative analyses in the next section; because SAM optimizer directly minimizes $\max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_L}(\theta+\epsilon)$. Hence, the importance of the second term in the upper bound, $\max_{\|\epsilon\|\leq\rho} L_{\mathscr{X}_U}(\theta+\epsilon)$, becomes more significant for Adam optimizer. Figure 12 of Appendix A.2 provides the test accuracy along the acquisition iterations, which shows SAAL achieves higher accuracy quicker than baselines (see Figure 12a, 12d, or 12g).

To demonstrate that SAAL is also scalable in a high-resolution dataset, we additionally perform three iterative experiments for Imagenet. Figure 2 shows that SAAL outperforms other baselines in every acquisition iteration.

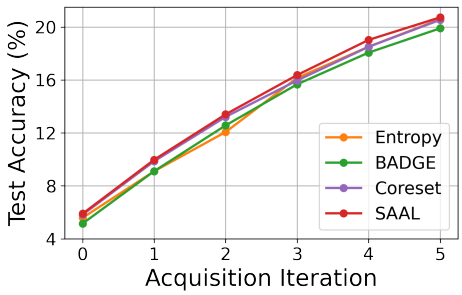


Figure 2. Comparison of test accuracy for ImageNet (%) using Adam optimizer.

Comparison of SAAL and SAM Our motivation started from minimizing the maximally perturbed loss bound in Eq. 6 for both labeled and unlabeled datasets. Having said that, SAM aims at minimizing the term w.r.t. the labeled dataset whereas SAAL aims at minimizing the term w.r.t. the unlabeled dataset. Hence, it should be noted that SAAL and SAM are orthogonal in their optimization to minimize

Table 1. Comparison of test accuracy (%) using Adam optimizer and SAM optimizer. The best performance is indicated as boldface, and we represent the second best performance as underline. (- represents that we failed to converge training when using SAM optimizer.)

	Fashion		SVHN		CIFAR-10		CIFAR-100	
Method	Adam	SAM	Adam	SAM	Adam	SAM	Adam	SAM
Random	81.2 ± 0.5	83.7 ± 0.3	72.4 ± 0.9	78.1 ± 1.1	50.7 ± 1.5	52.6 ± 2.8	43.3 ± 0.3	44.0 ± 0.7
Entropy	81.5 ± 1.4	84.1 ± 0.2	73.1 ± 1.0	77.5 ± 3.2	51.9 ± 1.8	54.6 ± 0.4	44.4 ± 0.7	44.1 ± 1.0
Coreset	83.8 ± 0.7	84.4 ± 0.6	75.3 ± 5.8	$\mathbf{78.9 \pm 1.3}$	51.7 ± 1.0	53.9 ± 1.3	44.4 ± 0.5	47.6 ± 1.4
LL4AL	83.5 ± 1.8	83.2 ± 1.4	75.1 ± 1.7	72.2 ± 0.2	51.7 ± 0.4	50.2 ± 1.1	43.9 ± 0.3	$35.7 \pm .01$
VAAL	83.4 ± 0.1	84.1 ± 0.6	73.4 ± 1.3	77.1 ± 0.8	52.0 ± 0.9	53.1 ± 0.9	44.8 ± 0.3	45.5 ± 0.4
BADGE	85.4 ± 0.6	86.2 ± 0.2	74.9 ± 1.1	78.8 ± 0.9	52.3 ± 2.2	56.8 ± 1.9	45.7 ± 0.6	47.4 ± 0.7
ProbCover	84.0 ± 0.2	-	74.3 ± 0.5	-	54.1 ± 0.6	-	42.6 ± 0.6	-
SAAL	85.6 ± 0.2	85.0 ± 0.3	76.5 ± 1.0	77.1 ± 1.0	52.3 ± 2.3	56.0 ± 1.2	46.6 ± 0.5	$\mathbf{48.4 \pm 0.9}$
w/ k-means++	$\mathbf{85.8 \pm 0.8}$	$\mathbf{86.3 \pm 0.5}$	$\mathbf{76.8 \pm 0.7}$	78.8 ± 1.0	$\mathbf{54.4 \pm 0.9}$	$\mathbf{57.0 \pm 1.1}$	$\mathbf{47.6 \pm 0.9}$	46.4 ± 0.1

the generalization gap of Eq. 6. We can infer the effect of SAAL and SAM, respectively, from Table 1. Taking CIFAR-10 dataset as an example, Random with Adam optimizer, which shows test accuracy of 50.7%, is the most naive baseline without any concerns on the minimization of the upper bound terms. Then, when we fix Random as the acquisition and turn the optimizer to SAM, it shows the test accuracy of 52.6%, whose gain is interpreted as the effect of SAM, i.e., the effect of minimizing the upper bound term w.r.t. labeled dataset. On the other hand, when we fix Adam as an optimizer and utilize the acquisition of SAAL, we achieve the test accuracy of 54.4% as the effect of SAAL, i.e., the effect of minimizing the upper bound w.r.t. unlabeled dataset. Finally, using both SAAL and SAM together shows the highest test accuracy of 57.0%, which convinces our motivation to minimize the upper bound of Eq.6.

Time Complexity We compare the time complexity of SAAL and baselines because SAAL has additional steps for finding the maximum perturbation over the acquisition calculations. We used CIFAR-10 and measured the time for a single iteration of acquisition and training. Figure 3 shows the wall-time by log scale. The results of Random acquisition show that the SAM optimizer takes twice longer time than the Adam optimizer, because it takes two steps of gradient calculation. However, the gap between Adam and SAM becomes smaller when using other active learning algorithms, indicating that the time for calculating acquisition score is the largest bottleneck. SAAL calculates the perturbation, ϵ, for every single unlabeled instance, instead of batch-wise calculation; so it takes longer than most of the other baselines. The time complexity of SAAL can be reduced if we adopt the improved SAM models (Du et al., 2021; 2022) that have been proposed for an efficient calculation. Additionally, Table 2 presents the trade-off between the wall-time and the batch size. Basically, we may increase the batch-size to reduce the calculation time of perturbation maximization, so this will provide the maximum perturbation to batch instances, not a single instance. This treatment drastically reduces the wall-time while maintaining performance improvement.

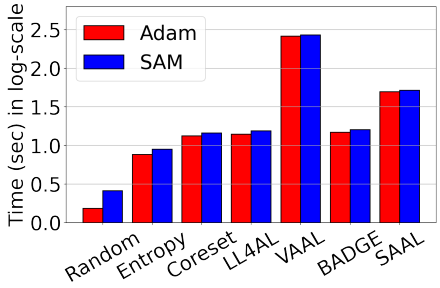


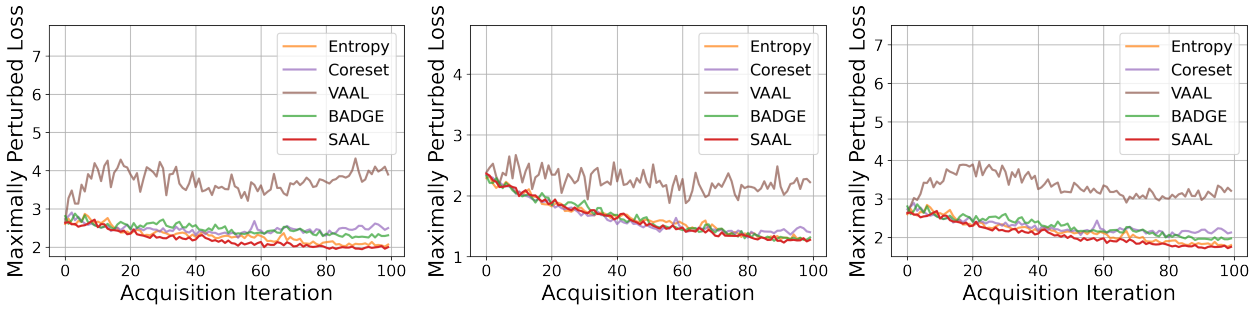
Figure 3. Comparison of time complexity.

Table 2. Test Accuracy on CIFAR-10 and Time Complexity of Batch-wise Perturbation.

Method	BS	Adam		SAM	
		Test accuracy	Time	Test accuracy	Time
BADGE	-	52.3 ± 2.2	14.8 s	56.8 ± 1.9	16.0 s
SAAL	1	54.4 ± 0.9	49.6 s	57.0 ± 1.1	51.7 s
	10	54.0 ± 1.0	16.8 s	57.7 ± 0.7	18.9 s
	100	53.6 ± 2.3	8.0 s	56.0 ± 1.5	10.3 s
	200	54.1 ± 1.3	7.5 s	56.2 ± 1.2	9.9 s

Qualitative Analysis Figure 4 supports the conjecture for the advantage of SAAL by anticipating the flat local minima in the acquisition process. Figure 4 measures the maximally perturbed loss for the labeled dataset, $\mathscr{X}_L$; the unlabeled dataset, $\mathscr{X}_U$; and the total dataset, $\mathscr{X}_L \cup \mathscr{X}_U$. We compare the results between the models trained with the SAM optimizer. Since it is computationally hard to calculate the corresponding perturbation for every single unlabeled instance, $x_u \in \mathscr{X}_U$, we uniformly sample 2,000 unlabeled instances from $\mathscr{X}_U$ at each iteration; and we report the averaged results for three independently repeated trials.

Figure 4a shows the maximally perturbed loss of $\mathscr{X}_U$ when using SAM optimizer. If we compare the result of SAAL with the results of baselines, SAAL shows the lowest value of the maximally perturbed loss, because SAAL selected the instances with high values of perturbed loss, and SAAL removed such instances by passing those instances to the



(a) Unlabeled dataset, $\mathscr{X}_U$ (b) Labeled dataset, $\mathscr{X}_L$ (c) Total dataset, $\mathscr{X}_U \cup \mathscr{X}_L$

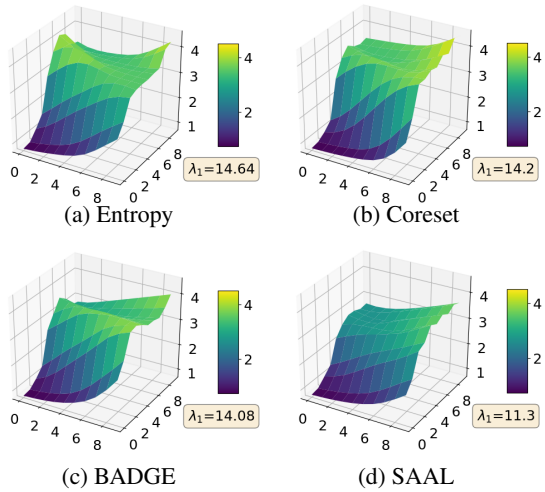
Figure 4. Maximally perturbed loss of CIFAR-10 during the active learning iterations, trained by SAM.

labeled dataset. Figure 4b shows the maximally perturbed loss of $\mathscr{X}_L$ when using SAM optimizer. This loss also indicates the flatness of the model; the lower value of the maximally perturbed loss of $\mathscr{X}_L$ indicates that the model does not change the result even if the parameter is changed in a small range, which refers to the flat model (Keskar et al., 2017; Neyshabur et al., 2017). Hence, SAAL results in a flat network compared to the baselines.

We conjecture that the flat model attained by SAAL is explained by the look-ahead concept (Roy & McCallum, 2001; Konyushkova et al., 2017; Kim et al., 2021). If we are planning to minimize $\max_{\|\epsilon\| \leq \rho} L_{\mathscr{X}}(\theta + \epsilon)$ by SAM optimizer, SAAL looks ahead the high values of the $\max_{\|\epsilon\| \leq \rho} L_{\mathscr{X}}(\theta + \epsilon)$ from unlabeled instances, and SAAL actively selects such unlabeled instances to flatten the future response surface.

Finally, Figure 4c shows the maximally perturbed loss of the total dataset, which is equivalent to the upper bound in Eq. 7. As confirmed in the figure, SAAL achieves the lowest upper bound, which indicates that the model trained with SAAL is more likely to achieve a lower population loss, which is our ultimate goal of minimization objective. When comparing the results of using SAM (Figure 4a - Figure 4c) and the results of using Adam (Figure 11a - Figure 11c in Appendix A.1), the gap between SAAL and other baselines becomes clearer in using the Adam optimizer.

Visualization of Loss Landscape SAAL aims at constructing a flat model by adaptively 1) selecting instances with high sharpness and 2) training the model to quickly decrease the loss for instances result in sharpness. Hence, we visualize the loss landscape with the first eigenvalue of loss Hessian matrix (Li et al., 2018). Appendix A.6 provides the detailed visualization formula and the full enumeration of figures. Figure 5 provides the loss landscape of SAAL and baselines, and the visual inspection and the first eigenvalue confirm that SAAL has a more flattened loss landscape.



(a) Entropy (b) Coreset (c) BADGE (d) SAAL

Figure 5. Loss landscapes for Fashion, optimized by Adam.

4.2. Ablation Study on Image Classification

Robustness to Class Imbalanced Figure 5 demonstrates that SAAL achieves a flat loss landscape, indicating a desirable property of the method. In order to further assess the robustness of SAAL, we conducted additional experiments and compared it with other baselines. Specifically, we created a long-tailed CIFAR-10 dataset and performed experiments under low-budget settings. Table 3 shows that SAAL effectively handles the imbalanced scenario, outperforming other baselines.

Table 3. Test Accruacy on long-tailed CIFAR-10 using Adam optimizer.

Method	Test Accuracy
Random	21.03 ± 0.89
Entropy	21.70 ± 0.95
Coreset	20.14 ± 1.23
BADGE	21.69 ± 1.00
SAAL	$\mathbf{23.03 \pm 1.11}$

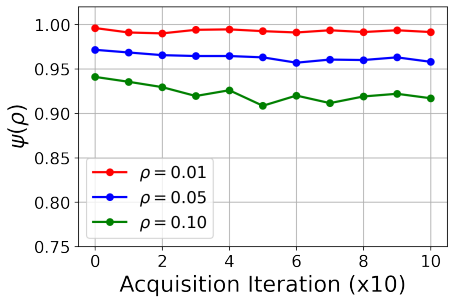


Figure 6. Proportion of unlabeled instances satisfying the assumption in Proposition 3.2, with varying ρ.

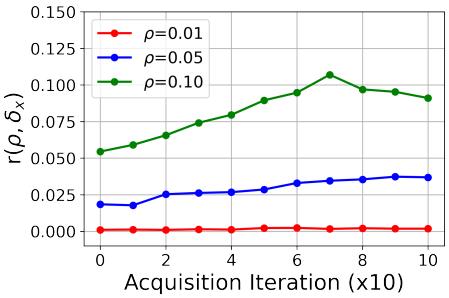


Figure 7. Averaged value of the margin, δ_x, in Theorem 3.1 for the unlabeled dataset; with varying ρ.

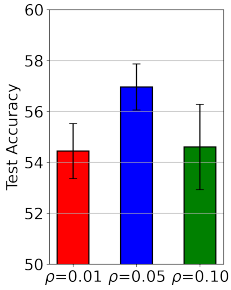


Figure 8. Test accuracy; with varying ρ.

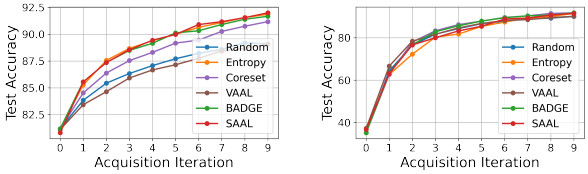
Sensitivity Analysis on ρ SAAL introduces a hyperparameter, ρ, which represents the size of the perturbation region, ϵ. Hence, we conduct the sensitivity analysis on ρ with the CIFAR-10 dataset, and we set the candidate values for ρ as 0.01, 0.05, and 0.10.

First, we examined the validity of Theorem 3.1 by investigating if the network with the maximally perturbed parameters keeps the predicted label as same as the original network. Figure 6 shows the proportion of unlabeled data instances whose predicted labels remain the same by the perturbed network during the active learning iterations; that is $\psi(\rho) := \frac{1}{|\mathscr{X}_U|} \sum_{x \in \mathscr{X}_U} \mathbf{1}_{\text{argmax}_j f_{\theta+\hat{\epsilon}}(x)_j = \hat{y}}$, where $\mathbf{1}_A$ is the indicator function. When the size, ρ, of the perturbation, ϵ, is zero (equivalently, if we do not perturb the network); then the inequality of Eq. 10 is satisfied for all instances, by the definition of the pseudo-label, $\hat{y}$. As we increase the value of ρ, some instances fail to keep the predicted label as the same as $\hat{y}$, because the parameter of the model changes drastically, so that the model loses the prediction ability that it has learned so far.

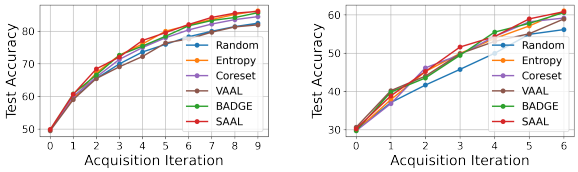
Also, we examined the validity of Proposition 3.2 by investigating the value of the margin, δ_x, for the unlabeled data instances in Figure 7. It should be noted that δ_x is not our hyperparameter, but a dependent variable subject to change by ρ. We only investigate δ_x to reveal the characteristics of ρ, not for the hyperparameter optimizations. To show how the value of the margin, δ_x, affects the inequality, we measure the relative value of the margin, δ_x, compared to the maximally perturbed loss, $\max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon)$; that is $r(\rho, \delta_x) := \frac{1}{|\mathscr{X}_U|} \sum_{x \in \mathscr{X}_U} \frac{\delta_x}{\max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon)}$. From the analyses of Figure 6 and 7, we adopted $\rho = 0.05$, because this value 1) keeps the predicted label of data instance from the original network with high probability and 2) keeps the value of the margin relatively small compared to the max perturbed loss w.r.t. the ground-truth label, while $\rho = 0.05$ is confirmed to perturb the parameters of the network effectively (Foret et al., 2020).

The proper selection of ρ also affects the test accuracy, as shown in Figure 8. If we select ρ with a too small value, that is $\rho = 0.01$, the parameter of the model is not perturbed enough to measure the sharpness, so SAAL cannot catch the informative instances. If we select ρ with a too-large value, that is $\rho = 0.10$, the maximally perturbed loss 1) does not satisfy Proposition 3.2, as confirmed in Figure 6, and 2) have too large value of margin, as confirmed in Figure 7. Meanwhile, a proper value of $\rho = 0.05$ for the perturbation, ϵ, shows the best performance.

Budget Variation To demonstrate SAAL is also scalable in high budget setting, we conduct an additional experiment. We follow the setting from (Yoo & Kweon, 2019); we increase the budget to 1,000 instances but decrease the iteration of acquisition to nine steps for Fashion, SVHN, and CIFAR-10. For CIFAR-100, we similarly increase initial labeled dataset to 5,000 but acquire the 2,500 unlabeled instances for six iterations. For further settings containing hyperparameters, we report in Appendix A.5. While Appendix A.3 shows figures from all cases, Figure 9 shows that SAAL is still the best result with a small margin.



(a) Fashion (b) SVHN



(c) CIFAR-10 (d) CIFAR-100

Figure 9. Test accuracy for high budget setting along the acquisition iteration; with SAM optimizers.

4.3. Object Detection

To show the effectiveness of SAAL in a complex task, we conduct an object detection task. Object detection returns the locations of semantic objects and the corresponding labels for a given input image, x. Hence, the loss for training detection model consists of the bounding box regression loss and the classification loss.

We experiment with PASCAL VOC 2007 and 2012 dataset (Everingham et al., 2010), which contains 5,011 images and 4,952 images with 20 object classes, respectively. We adopt Single Shot Multibox Detector (SSD) (Liu et al., 2016) as the detection model. To apply SAAL for object detection, we perturb the parameters to maximize the classification loss; and use the summation of the perturbed loss from every corresponding detection box in the image, x, as the acquisition score for x. Afterward, we select the images with the highest scores. We construct the initial labeled dataset with 1,000 randomly selected images, and we select additional 1,000 instances at every acquisition iterations, so that we attain 10,000 final instances with nine repeated acquisitions. We train the model for 300 epochs with a batch size of 32. Figure 10 reports the mean average precision (mAP) for three repeated trials of SAAL and baselines. As shown in the figure, SAAL achieves high performance at the earlier iterations and shows the highest mAP of 0.7541 at the last iteration; while BADGE, Entropy, and Random show 0.7493, 0.7518, and 0.7403, respectively.

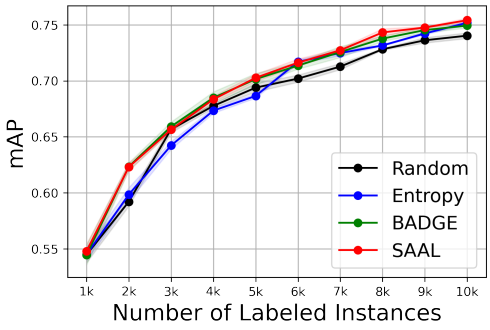


Figure 10. mAP of object detection task with PASCAL VOC 2007+2012.

4.4. Domain Adaptive Semantic Segmentation

Recently, active learning strategy is utilized for domain adaptive semantic segmentation (Xie et al., 2022). We experiment the semantic segmentation from a source domain SYNTHIA (Ros et al., 2016) to a target domain CityScapes (Cordts et al., 2016). The base strategy of acquisition follows "Region-based Annotating", which queries a pixel-wise but acquires the neighborhood of a high-scored pixel (Xie et al., 2022). RIPU (Xie et al., 2022) calculates the acquisition score per pixel, which consists of a multiplication of diversity and uncertainty score on a pixel. We similarly calculate the acquisition score (RI-SAAL) by multiplying the diversity score from (Xie et al., 2022) with the acquisition score from SAAL (see details in Appendix A.8). Table 4 confirms that RI-SAAL outperforms other baselines.

Table 4. mIOU of domain adaptive semantic segmentation from SYNTHIA to CityScapes. The best performance is indicated as boldface.

Method	mIOU
Random	68.3
Entropy	68.6
RIPU (Xie et al., 2022)	70.2
RI-SAAL	**70.6**

5. Conclusion and Future Works

We propose a new active learning method named Sharpness-Aware Active Learning, or SAAL. The proposed method considers the loss sharpness of data instances, which is strongly related to the generalization performance of deep learning. Furthermore, we derive the upper bound of SAAL acquisition score and find the connection to the recent active learning methods; as well as the connection to the first eigenvalue of loss Hessian matrix, which is widely used as the indicator of loss sharpness. In various experiments with benchmark datasets, SAAL shows better performance than baselines.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C200981612). Also, this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (NO. 2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data).

References

Angluin, D. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

Angluin, D. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020.

Atlas, L., Cohn, D., and Ladner, R. Training connectionist networks with queries and selective sampling. *Advances in neural information processing systems*, 2, 1989.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ben-David, S., Lu, T., and Pál, D. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pp. 33–44, 2008.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220, 2010.

Cohn, D., Atlas, L., and Ladner, R. Improving generalization with active learning. *Machine learning*, 15(2): 201–221, 1994.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Dasgupta, S. and Hsu, D. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 208–215, 2008.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2021.

Du, J., Zhou, D., Feng, J., Tan, V. Y., and Zhou, J. T. Sharpness-aware training for free. *arXiv preprint arXiv:2205.14083*, 2022.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2):303–338, 2010.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

Freeman, L. *Elementary applied statistics: for students in behavioral science*. Wiley, 1965. URL https://books.google.co.kr/books?id=r4VRAAAAMAAJ.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Kaur, S., Cohen, J., and Lipton, Z. C. On the maximum hessian eigenvalue and generalization. *arXiv preprint arXiv:2206.10654*, 2022.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., and Smelyanskiy, M. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Kim, Y.-Y., Song, K., Jang, J., and Moon, I.-C. Lada: Look-ahead data acquisition via augmentation for deep active learning. *Advances in Neural Information Processing Systems*, 34:22919–22930, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Konyushkova, K., Sznitman, R., and Fua, P. Learning active learning from data. *arXiv preprint arXiv:1703.03365*, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *SIGIR'94*, pp. 3–12. Springer, 1994.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6391–6401, Red Hook, NY, USA, 2018. Curran Associates Inc.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3234–3243, 2016. doi: 10.1109/CVPR.2016.352.

Roy, N. and McCallum, A. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

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

Settles, B., Craven, M., and Ray, S. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

Sinha, S., Ebrahimi, S., and Darrell, T. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5972–5981, 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xie, B., Yuan, L., Li, S., Liu, C. H., and Cheng, X. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8068–8078, June 2022.

Yehuda, O., Dekel, A., Hacohen, G., and Weinshall, D. Active Learning Through a Covering Lens. *arXiv preprint arXiv:2205.11320*, 2022.

Yoo, D. and Kweon, I. S. Learning loss for active learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2019.

Zhang, G., Wang, C., Xu, B., and Grosse, R. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018a.

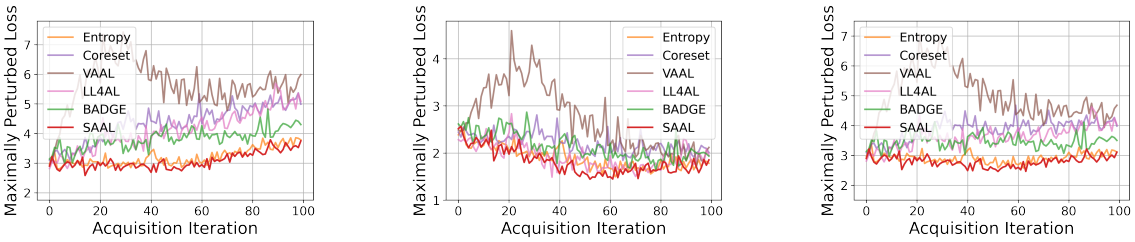
Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018b.

Zhu, X. J. Semi-supervised learning literature survey. 2005.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N., Tatikonda, S., Duncan, J., and Liu, T. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022.

A. Appendix

A.1. Maximally perturbed loss with Adam optimizer



(a) Maximally perturbed loss of $\mathscr{X}_U$ with Adam

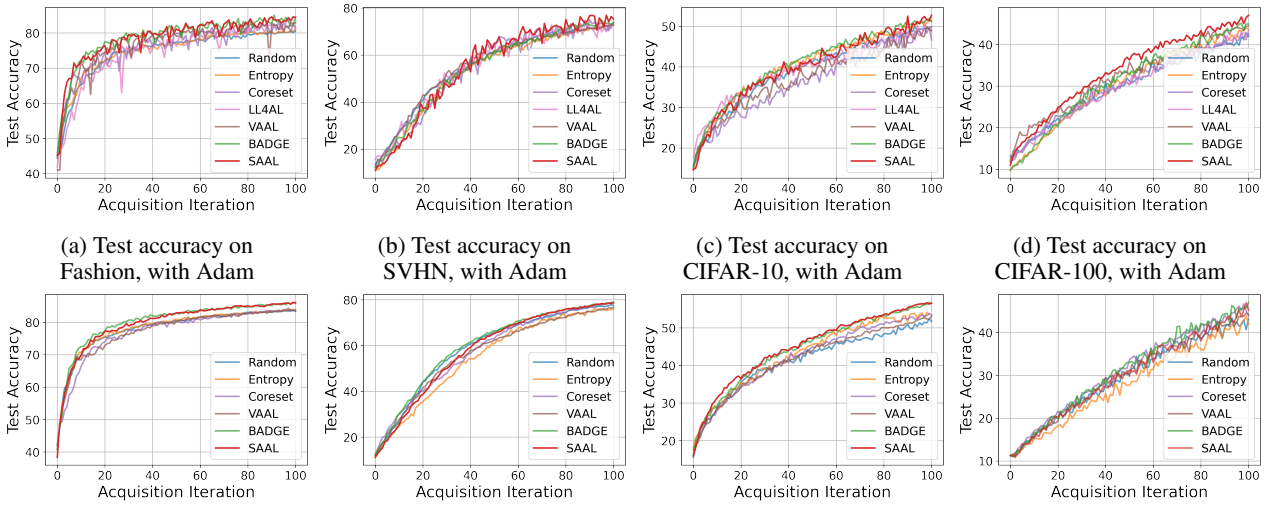
(b) Maximally perturbed loss of $\mathscr{X}_L$ with Adam

(c) Maximally perturbed loss of $\mathscr{X}_L \cup \mathscr{X}_U$ with Adam

Figure 11. Maximally perturbed loss of the labeled dataset, unlabeled dataset, and total dataset during the active learning iterations. (a) - (c) are the results of the model trained by Adam optimizer.

A.2. Test accuracy of classification for low budget setting

For low budget setting, we provide the learning curve of SAAL and baselines along the acquisition iterations.



(a) Test accuracy on Fashion, with Adam

(b) Test accuracy on SVHN, with Adam

(c) Test accuracy on CIFAR-10, with Adam

(d) Test accuracy on CIFAR-100, with Adam

(e) Test accuracy on Fashion, with SAM

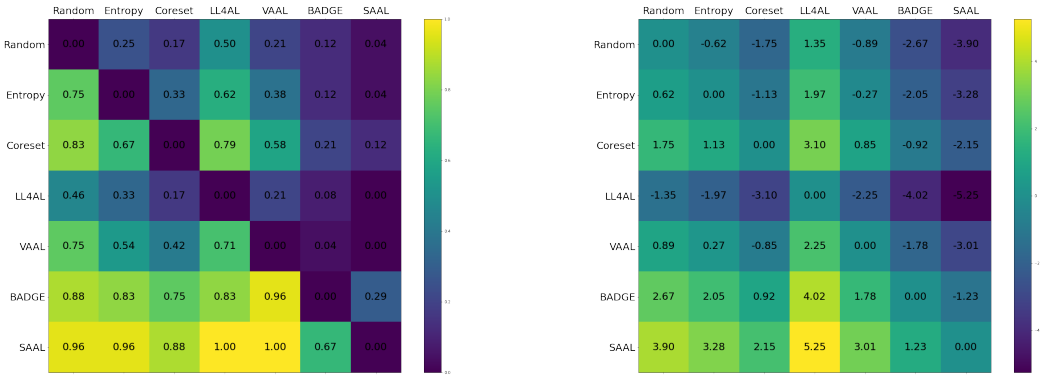
(f) Test accuracy on SVHN, with SAM

(g) Test accuracy on CIFAR-10, with SAM

(h) Test accuracy on CIFAR-100, with SAM

Figure 12. Test accuracy for low budget setting along the acquisition iteration; with Adam and SAM optimizers.

To show the improvement of SAAL, we also provide the overall comparison in Figure 13. We consider all the N comparison cases, where N contains the number of random seeds and the number of datasets, and the number of optimizers. Figure 13a shows the pairwise comparison, where $(i, j)^{th}$ cell indicates the proportion of the number when the i^{th} algorithm beats the j^{th} algorithm. Figure 13b shows the pairwise comparison, where $(i, j)^{th}$ cell indicates the averaged value of the performance gain achieved by the i^{th} algorithm compared to the j^{th} algorithm. The figures prove that SAAL outperforms the baselines in most cases.



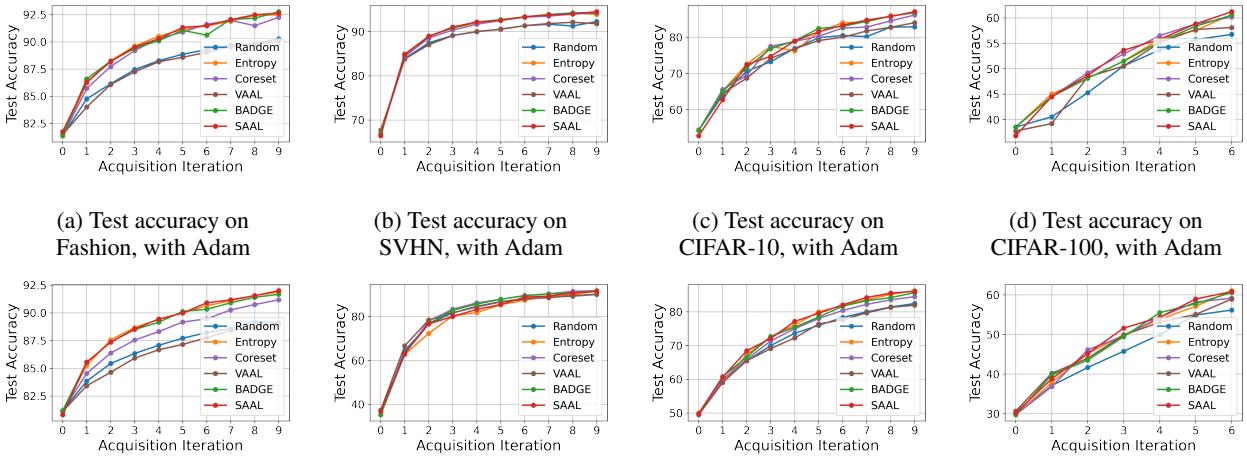
(a) Overall comparison on number of beating trials

(b) Overall comparison on performance gain

Figure 13. Overall comparison of SAAL with baselines.

A.3. Test accuracy of classification for high budget setting

For high budget setting, we provide the learning curve of SAAL and baselines along the acquisition iterations.



(a) Test accuracy on Fashion, with Adam

(b) Test accuracy on SVHN, with Adam

(c) Test accuracy on CIFAR-10, with Adam

(d) Test accuracy on CIFAR-100, with Adam

(e) Test accuracy on Fashion, with SAM

(f) Test accuracy on SVHN, with SAM

(g) Test accuracy on CIFAR-10, with SAM

(h) Test accuracy on CIFAR-100, with SAM

Figure 14. Test accuracy for high budget setting along the acquisition iteration; with Adam and SAM optimizers.

A.4. Details of experiment for low budget setting

For Fashion, SVHN, and CIFAR-10, we construct the initial labeled dataset with 20 instances, which are random but balanced; and we select 10 additional instances with the highest acquisition score among the randomly selected 2,000 unlabeled instances per each iteration. For CIFAR-100, the initial labeled dataset consists of 1,000 instances, and we select 100 additional instances for 100 repeated iterations. For ImageNet, the initial labeled dataset consists of 5,000 instances, and we select 5,000 additional instances for five repeated iterations. Here, SAAL introduces the perturbation size, ρ, of the perturbation, ϵ, in Eq. 8, and we set the value of ρ as 0.05 for all the datasets.

A.5. Details of experiment for high budget setting

We adopt Resnet-18 as backbone of our classifier. We train the network for 200 epochs after each acquisition step, using Adam optimizer with a learning rate of 0.0005; and SAM optimizer with a learning rate of 0.001 for Fashion, SVHN, CIFAR-10 and 0.1 for CIFAR-100. In high budget setting, we additionally optimize ρ of SAAL; Table 5 shows details of finding ρ.

Table 5. Optimizing the proper ρ for SAAL in high budget setting.

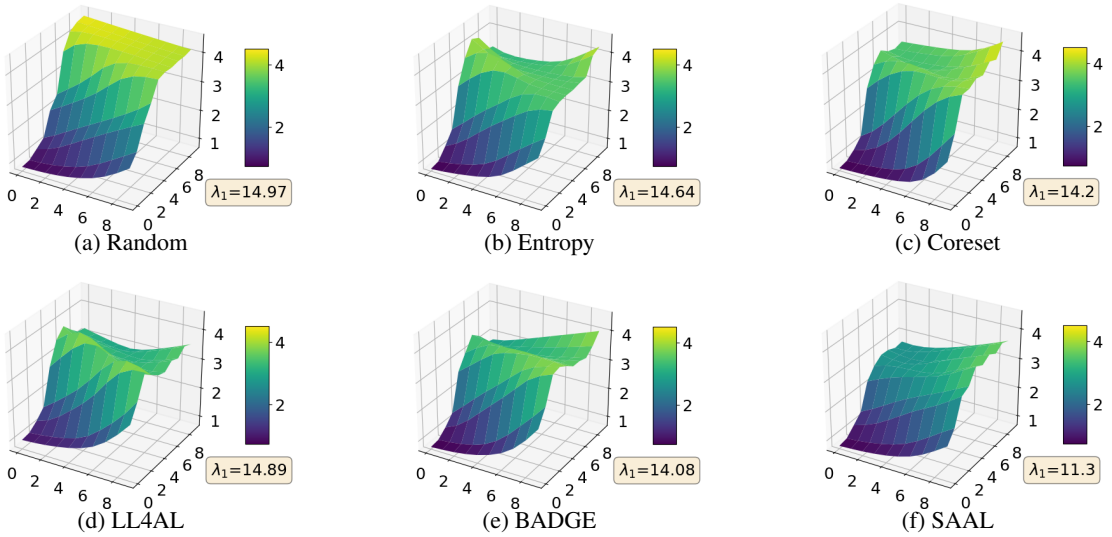
Dataset	Adam optimizer	SAM optimizer
Fashion	{0.03,**0.04**,0.05,0.06}	{0.03,**0.04**,0.05,0.06}
SVHN	{0.03,**0.04**,0.05,0.06}	{**0.03**,0.04,0.05,0.06}
CIFAR-10	{**0.03**,0.04,0.05,0.06}	{**0.03**,0.04,0.05,0.06}
CIFAR-100	{**0.03**,0.04,0.05,0.06}	{0.03,**0.04**,0.05,0.06}

A.6. Loss landscape

When moving the weight θ along random directions d_1 and d_2 with magnitude α and β, plotting the loss change as the below:

$$g(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_{\theta+\alpha d_1+\beta d_2}(x), y) \tag{14}$$

For a fair comparison, we calculate the loss of Fashion training set and perturb the loss to the same random directions.



(a) Random (b) Entropy (c) Coreset

(d) LL4AL (e) BADGE (f) SAAL

Figure 15. (a)-(f) are loss landscape for Fashion, optimized by Adam optimizer.

A.7. Additional experiments

A.7.1. Details of experiment for class-imbalanced setting

We followed low budget setting of CIFAR-10 in Appendix A.4, and long-tailed CIFAR-10 is composed with the samples by exponentially imbalanced class ratios.

A.7.2. Ablation study with k-means++

To compare the performance of SAAL when applying k-means++ with uncertainty-based active learning methods, we conducted an additional ablation study. Table 6 indicates that applying k-means++ algorithm improves baselines by considering diversity, but SAAL still outperforms the other methods.

Table 6. Test Accruacy of uncertainty-based active learning methods with k-means ++ on CIFAR-10.

Method	Adam	SAM
Entropy w/k-means++	51.3 ± 0.3	54.9 ± 1.1
LL4AL w/k-means++	52.7 ± 1.2	55.3 ± 1.1
SAAL w/k-means++	$\mathbf{54.4 \pm 0.9}$	$\mathbf{57.0 \pm 1.1}$

A.7.3. Correlations of acquisition scores and upper bounds in Theorem 3.3

We compare the correlation between other method's acquisition score and upper bound terms in Theorem 3.3. Table 7 indicates that SAAL and BADGE have a high correlation with upper bounds, where the BADGE has a higher correlation of 1^{st} Eigenvalue of loss Hessian matrix.

Table 7. Correlation value between acquisition scores and upper bound terms on CIFAR-10. Since BADGE use the gradient norm as acquisition score, we mark it -.

Method	Task loss	Gradient norm	1^{st} Eigenvalue of loss Hessian matrix
Entropy	0.939	0.917	0.885
BADGE	0.961	–	**0.937**
SAAL	**0.988**	**0.976**	0.924

In addition, we compare the correlation value among upper bound terms, and we conduct the experiment about how many same data points are selected. In Table 8, we confirm that applying SAAL shows high positive correlation with all the upper bound terms, while other upper bound terms do not. This indicates that, for example, the selected instances with loss as acquisition function are not assured to have high eigenvalue compared to SAAL as acquisition function.

Table 8. Correlation value among upper bound terms on CIFAR-10.

	Task loss	Gradient norm	1^{st} Eigenvalue of loss Hessian matrix
Task loss	–	0.961	0.863
Gradient norm	0.961	–	0.937
1^{st} Eigenvalue of loss Hessian matrix	0.863	0.937	–

Table 9 indicates that the proportion for intersection of selected instances is similar to the tendency of Figure 1a.

Table 9. Proportion of selecting the same data points per number of selections, k.

k	20	40	60	80	100
Task loss	0.600	0.825	0.918	0.938	0.964
Gradient norm	0.350	0.742	0.827	0.897	0.938
1^{st} Eigenvalue of loss Hessian matrix	0.100	0.442	0.650	0.788	0.881

A.7.4. Effect of low sharpness instances

We conduct an experiment to select instances with low sharpness, and Table 10 indicates that the acquisition of low sharpness degrades test accuracy. To analyze the degradation of the performance, we confirmed that the selected instances had a very low value of the acquisition score, $\max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon) \approx 0$. With these instances, the updated upper bound term w.r.t the labeled dataset, i.e., $\pi_L \max_{\|\epsilon\| \leq \rho} L_{\mathscr{X}_L}(\theta + \epsilon)$ will be merely changed. This indicates that the model parameter, θ, is not updated with active learning, and consequently shows a very low test accuracy.

Table 10. Comparison of test accuracy (%) between low sharpness and high sharpness acquisition on CIFAR-10.

Optimizer	Adam	SAM
SAAL-Reverse	36.9 ± 1.1	28.5 ± 0.2
SAAL	$\mathbf{54.4 \pm 0.9}$	$\mathbf{57.0 \pm 1.1}$

A.8. Details of Domain Adaptive Semantic Segmentation

In (Xie et al., 2022), the acquisition score is calculated as a multiplication of the diversity score (Region Impurity) and uncertainty score (Prediction Uncertainty). Instead of Prediction Uncertainty, we utilize the score from SAAL. Then, the classifier may select high-sharpness valued pixels, whose neighborhoods contain diverse classes

Region impurity measures how neighbors of a pixel contain various classes. First, we define the neighborhood set of a pixel (i, j) as follows:

$$N_k(i,j) = \{(u,v)|\, |u-i| \leq k, |v-j| \leq k\}$$

Pseudo-label $\hat{Y}^{(i,j)}$ is utilized to divide subset of a pixel (i, j) and Region impurity $P^{(i,j)}$ is calculated as follows:

$$N_k^c(i,j) = \{(u,v) \in N_k(i,j)|\, \hat{Y}^{(u,v)} = c\}$$

$$P^{(i,j)} = -\sum_{c=1}^{C} \frac{|N_k^c(i,j)|}{|N_k(i,j)|} \log \frac{|N_k^c(i,j)|}{|N_k(i,j)|}$$

We can define the pixel-wise score from SAAL as $S^{(i,j)} = f_{acq}^{(i,j)}$. Finally, we utilize the final acquisition function $A^{(i,j)} = P^{(i,j)} S^{(i,j)}$.

A.9. Proof Details

A.9.1. Proof of Theorem 3.1

Theorem A.1. *For a data instance x, let $\hat{y}$ be the pseudo-label predicted by the network f_θ and $\bar{y}$ be the ground-truth label. Then, the maximally perturbed loss calculated with $(x, \hat{y})$ is a lower bound of the maximally perturbed loss calculated with $(x, \bar{y})$; with a non-negative margin, δ_x, as the below:*

$$\max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon) \leq \max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon) + \delta_x.$$

Proof. The cross-entropy loss, $l(x, y; \theta)$, is represented with the logit vector $f_\theta(x) \in \mathbb{R}^{|\mathscr{Y}|}$ as the below:

$$\begin{aligned} l(x,y;\theta) &= -\ln \frac{\exp(f_\theta(x)_y)}{\sum_j \exp(f_\theta(x)_j)} \\ &= -\ln\left(\exp(f_\theta(x)_y)\right) + \ln \sum_j \exp(f_\theta(x)_j) \\ &= \ln \sum_j \exp(f_\theta(x)_j) - f_\theta(x)_y. \end{aligned}$$

Then, the maximally perturbed loss of a data pair (x, y) is represented as the below:

$$\max_{\|\epsilon\| \leq \rho} l(x, y; \theta + \epsilon) = \max_{\|\epsilon\| \leq \rho} (\ln \sum_j \exp(f_{\theta+\epsilon}(x)_j) - f_{\theta+\epsilon}(x)_y).$$

Since the pseudo-label, $\hat{y}$, satisfies $\hat{y} = \text{argmax}_{j \in \mathscr{Y}} f_\theta(x)_j$ by the definition, it holds that $f_\theta(x)_{\hat{y}} \geq f_\theta(x)_j$ for all $j \in \mathscr{Y}$. Let $\hat{\epsilon} = \text{argmax}_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon)$. Define the margin, δ_x, as $\delta_x := [\max_j \{f_{\theta+\hat{\epsilon}}(x)_j - f_{\theta+\hat{\epsilon}}(x)_{\hat{y}}\}]_+$ where $[\cdot]_+ = \max\{\cdot, 0\}$. Then, the following holds.

$$\begin{aligned} \max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon) &= \ln \sum_j \exp(f_{\theta+\hat{\epsilon}}(x)_j) - f_{\theta+\hat{\epsilon}}(x)_{\hat{y}} \\ &\leq \ln \sum_j \exp(f_{\theta+\hat{\epsilon}}(x)_j) - f_{\theta+\hat{\epsilon}}(x)_{\bar{y}} + \delta_x \\ &\leq \max_{\|\epsilon\| \leq \rho} \left(\ln \sum_j \exp(f_{\theta+\epsilon}(x)_j) - f_{\theta+\epsilon}(x)_{\bar{y}} \right) + \delta_x \\ &= \max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon) + \delta_x \end{aligned}$$

□

A.9.2. PROOF OF PROPOSITION 3.2

Proposition A.2. *For a data instance x and the corresponding pseudo-label $\hat{y}$, let $\hat{\epsilon}$ be the maximal perturbation over the parameters w.r.t. the loss $l(x, \hat{y}; \theta + \epsilon)$. If the perturbed network, $f_{\theta+\hat{\epsilon}}$, keeps the predicted label as the same as the label predicted from the original network, f_θ; then the maximally perturbed loss calculated with $(x, \hat{y})$ is a lower bound of the maximally perturbed loss calculated with $(x, \bar{y})$, as the below:*

$$\max_{\|\epsilon\| \leq \rho} l(x, \hat{y}; \theta + \epsilon) \leq \max_{\|\epsilon\| \leq \rho} l(x, \bar{y}; \theta + \epsilon).$$

Proof. Since the perturbed network, $f_{\theta+\hat{\epsilon}}$, keeps the predicted label as the same as the label predicted from the original network, f_θ; it holds that $\text{argmax}\, f_{\theta+\hat{\epsilon}}(x) = \text{argmax}\, f_\theta(x) = \hat{y}$ and accordingly $f_{\theta+\hat{\epsilon}}(x)_j \leq f_{\theta+\hat{\epsilon}}(x)_{\hat{y}}$ for all j. Hence, $\max_j \{f_{\theta+\hat{\epsilon}}(x)_j - f_{\theta+\hat{\epsilon}}(x)_{\hat{y}}\} \leq 0$. Thus, by the definition of the margin in Theorem 3.1, δ_x becomes zero.

□

A.9.3. PROOF OF THEOREM 3.3

Theorem A.3. *The acquisition function, f_{acq}^{SAAL}, of Eq. 8 is upper bounded by $l(\theta) + \rho\|\nabla_\theta l(\theta)\|_2 + \frac{1}{2}\rho^2\lambda_1 + \max_{\|v\|\leq 1} O(\rho^2 v^3)$; where $l(\theta)$ abbreviates the loss of a data pair, (x, y), and λ_1 is the first eigenvalue of the loss Hessian matrix.*

Proof. Recall that our acquisition function is $f_{acq}^{SAAL} = \max_{\|\epsilon\|\leq\rho} l(x_u, \hat{y}_u; \theta+\epsilon)$. Since we limit the size of the perturbation as $\|\epsilon\| \leq \rho$, we can write $\epsilon = \rho v$ with $\|v\| \leq 1$, and $\max_{\|\epsilon\|\leq\rho} l(x_u, \hat{y}_u; \theta + \epsilon) = \max_{\|\rho v\|\leq\rho} l(x_u, \hat{y}_u; \theta + \rho v) = \max_{\|v\|\leq 1} l(x_u, \hat{y}_u; \theta + \rho v)$. Then, by Taylor expansion of $l(x_u, \hat{y}_u; \theta + \rho v)$ w.r.t. θ, the below holds, where we abbreviate $l(x_u, \hat{y}_u; \theta)$ as $l(\theta)$.

$$\begin{aligned}
f_{acq}^{SAAL}(x_u; f_\theta) &= \max_{\|\epsilon\|\leq\rho} l(\theta + \epsilon) = \max_{\|v\|\leq 1} l(\theta + \rho v) \\
&= \max_{\|v\|\leq 1} \{l(\theta) + (\rho v)^T \nabla_\theta l(\theta) + \frac{1}{2}(\rho v)^T \nabla_\theta^2 l(\theta)(\rho v) + O((\rho v)^3)\} \\
&= l(\theta) + \max_{\|v\|\leq 1} \{(\rho v)^T \nabla_\theta l(\theta) + \frac{1}{2}(\rho v)^T \nabla_\theta^2 l(\theta)(\rho v) + O((\rho v)^3)\} \\
&\leq l(\theta) + \max_{\|v\|\leq 1} (\rho v)^T \nabla_\theta l(\theta) + \max_{\|v\|\leq 1} \frac{1}{2}(\rho v)^T \nabla_\theta^2 l(\theta)(\rho v) + \max_{\|v\|\leq 1} O((\rho v)^3) \\
&= \underbrace{l(\theta)}_{\text{Loss}} + \underbrace{\rho\|\nabla_\theta l(\theta)\|_2}_{\text{Gradient Norm}} + \underbrace{\frac{1}{2}\rho^2\lambda_1}_{1^{st}\text{ Eigenvalue}} + \max_{\|v\|\leq 1} O((\rho v)^3)
\end{aligned}$$

□