Debiased Active Learning with Variational Gradient Rectifier

Weiguo Chen[1*], Changjian Wang[1*†], Shijun Li[1], Kele Xu[1†], Yanru Bai[2], Wei Chen[1], Shanshan Li[1]

[1]College of Computer Science and Technology, National University of Defense Technology
[2]Academy of Medical Engineering and Translational Medicine, Tianjin University
{chenweiguo, wangcj, lishijun, chenwei}@nudt.edu.cn, xukelele@163.com, yr56_bai@tju.edu.cn

Abstract

The strategy of selecting "most informative" hard samples in active learning has proven a boon for alleviating the challenges of few-shot learning and costly data annotation in deep learning. However, this very preference towards hard samples engenders bias issues, thereby impeding the full potential of active learning. It has witnessed an increasing trend to mitigate this stubborn problem, yet most neglect the quantification of bias itself and the direct rectification of dynamically evolving biases. Revisiting the bias issue, this paper presents an active learning approach based on the Variational Gradient Rectifier (VaGeRy). First, we employ variational methods to quantify bias at the level of latent state representations. Then, harnessing historical training dynamics, we introduce Uncertainty Consistency Regularization and Fluctuation Restriction, which asynchronously iterate to rectify gradient backpropagation. Extensive experiments demonstrate that our proposed methodology effectively counteracts bias phenomena in a majority of active learning scenarios.

Introduction

The substantial success of many deep networks across diverse tasks is largely attributed to the availability of vast amounts of labeled data, yet this comes at a cost of annotation proportional to both the volume of data and the intricacy of labeling. Active learning has long been devoted to realizing the fastest possible convergence rate of deep networks with minimal labeling budget, thereby attaining the desired generalization performance (Liu et al. 2022; Zhang, Strubell, and Hovy 2022; Wan et al. 2023).

The objective of active learning strategies lies in the selective querying of "most informative" samples within a human-in-loop (Mosqueira-Rey et al. 2023) iterative framework. The quantification of "most informative" currently encompasses uncertainty-aware (Cao et al. 2019; Pleiss et al. 2020; Roth and Small 2006; Wu, Chen, and Huang 2022), variety-aware (Guo 2010), representation-aware (Sener and Savarese 2018), and predicting-auxiliary or synthesizing (Kye et al. 2023; Sinha, Ebrahimi, and Darrell 2019; Yoo and Kweon 2019) approaches. However, the combination

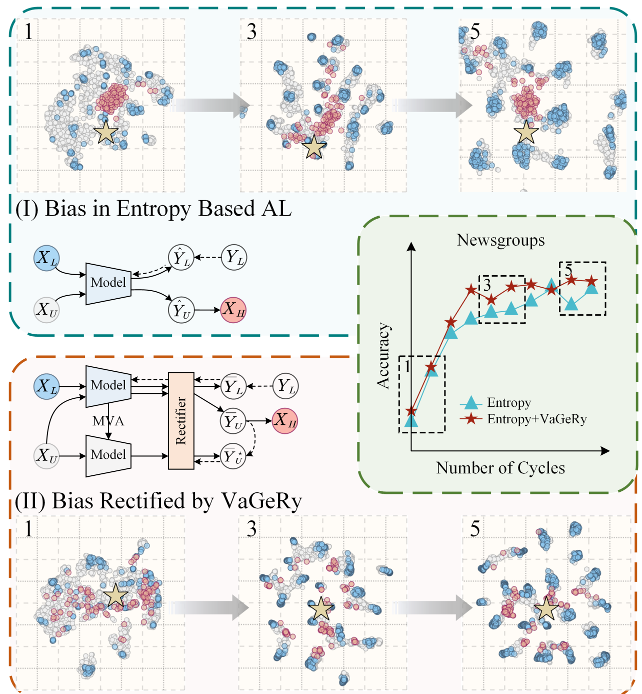

Figure 1: Comparison of bias phenomenon of the 1th, 5th and 9th iterations of (I) Entropy based active learning with bias phenomenon and (II) VeGeRy with rectified bias, corresponding to the three instances highlighted within boxes on the accuracy curves on Newsgroup dataset. The two intuitive frameworks in the blue and red dashed boxes respectively denote standard entropy based active learning and its VaGeRy-enhanced counterpart. The gray, blue, and red points in the scatter diagrams represent the unlabeled (X_U), labeled (X_L) samples and hard (X_H) samples respectively of a current iteration. The stars represent the approximated posterior distribution. It is obvious that the hard samples in (I) are clustered and the approximated posterior distribution has a large shift between iterations and behaves unstable. On the contotary, the hard samples in (II) are more evenly distributed and the approximated posterior distribution is rubust with a small shift between iterations.

[*]These authors contributed equally.
[†]Corresponding author

of the inherent limitations imposed by small sample sizes during the early stages of active learning iterations, coupled with the persistent inclination towards hard samples, engenders a bias phenomenon (Farquhar, Gal, and Rainforth 2021; Krishnan et al. 2021; Li et al. 2023; Chen et al. 2022) aptly characterized as the "Matthew Effect" by (Chen et al. 2022). Intuitively, model fitting to a limited number of annotated samples engenders *overfitting bias*, which, in turn, prompts active learning heuristics to selectively seek out hard examples, thereby introducing an appreciable degree of extremity in the distributional divergence between the training and overall population data, effectively constituting *statistical bias* (Farquhar, Gal, and Rainforth 2021). This bias issue introduces instability in training and slows down convergence rates, ultimately impeding the efficacy of active learning, as shown in Fig. 1. Hence, the measurement and mitigation of such bias concerns assume paramount importance.

Current remedies can be categorized into three broad classes: sampling optimization strategies, semi-supervised strategies and latent space representation strategies. Sampling optimization strategies mitigate bias by refining the hard sample selection mechanism, encompassing methods such as universal model data sampling (Xie et al. 2023), confidence-based pseudo-labeling (Elezi et al. 2022), and approaches rooted in training dynamics (Swayamdipta et al. 2020; Yao et al. 2022; Kye et al. 2023). Semi-supervised strategies incorporate consistency constraints on unlabeled data into the active learning training process, thereby enhancing model generalization across the broader data landscape and indirectly attenuating bias (Gao et al. 2020; Laine and Aila 2017; Verma et al. 2022; Athiwaratkun et al. 2019; Huang et al. 2024; Laine and Aila 2017). Latent space representation strategies endeavor to embed both labeled and unlabeled data into a shared latent space in an adversarial manner (Kim et al. 2021; Liu et al. 2023; Wu et al. 2023; Zhang et al. 2020), thereby enhancing the model's generalization capacity.

While the aforementioned methods have demonstrated their efficacy, they either exclusively revolve around sample selection heuristics or indiscriminately apply consistency constraints to stochastic gradient updates. In fact, bias manifests as a dynamic phenomenon that evolves alongside the iterative process of active learning. The majority of these approaches fail to directly confront bias itself, neglecting its quantification and monitoring. This prompts us to pose two critical inquiries: Can bias be effectively measured? And, can dynamically evolving bias be rectified?

In this paper, we concentrate on improving uncertainty-aware active learning methodologies without proposing specific enhancements to the underlying uncertainty measurement techniques. Instead, we introduce a rectification framework applicable to any chosen uncertainty metric, designed to mitigate bias arising during iterative processes under that metric. Specifically, we put forth an active learning framework with **Variational Gradient Rectifier (VaGeRy)**, which embeds both labeled and unlabeled data into a latent space, measures bias based on shared latent variables, and subsequently employs Uncertainty Consistency Regulation (UCR) along with Fluctuation Restriction (FR) constraints on unlabeled data, utilizing asynchronous iterations that draw upon historical training dynamics. In summary, the contributions are as follows:

- We revisit the issue of bias and, utilizing visualization techniques, assess its potential to impede active learning performance. Moreover, we explore a "quantification-to-rectification" pathway to alleviate bias phenomena.
- We propose an asynchronous alternating iterative framework, named Variational Gradient Rectifier, tailored to rectify biases in a diverse spectrum of uncertainty-aware methods, ultimately unlocking the untapped potential of active learning.
- We utilize latent state representations to measure bias phenomena and also apply training dynamics to the bias rectification process in active learning, integrating training dynamics with the design of UCR and FR.
- We conduct extensive experiments using VaGeRy to rectify six types of uncertainty measurements and two state-of-the-art methods across ten text datasets and two image datasets, demonstrating the effectiveness of VaGeRy's rectification capabilities.

Preliminaries

Let $(x_L, y_L) \in (X_L, Y_L)$ be a labeled sample pair from X_L and Y_L; (X_L^0, Y_L^0) are the initial labeled samples and labels; X_U denotes the pool of unlabeled samples; M_θ^0 is the randomly initialized model with parameters θ, and M_θ^i represents the model after the i-th iteration. Active learning is a Human-in-the-Loop process (Mosqueira-Rey et al. 2023) where "informative" samples from the unlabeled pool X_U—determined by uncertainty, representativeness, and variety—are selected based on the model's current and past versions $\{M_\theta^j\}_{j=0}^{i-1}$. After human labeling, these samples refresh the labeled pool X_L^{i-1}. Subsequent training with X_L^{i-1} updates the model to M_θ^i.

Bias and Debias in Active Learning

Due to the training data no longer following the population distribution, active learning inherently introduces biases (Krishnan et al. 2021; Parmar et al. 2023). Literature (Farquhar, Gal, and Rainforth 2021) highlights the inherent statistical biases and overfitting biases within active learning. An increasing body of research is now focused on how to mitigate harmful biases in this domain.

Training Dynamics significantly impact data diagnosis and selection (Swayamdipta et al. 2020). In dynamic consistency regularization, the temporal ensembling method (Laine and Aila 2017) integrates historical outputs for unlabeled data and substitutes them for the current model input, applying historical information in active learning. Models like mean teacher (Tarvainen and Valpola 2017) and COD (Huang et al. 2024, 2021) use exponential moving average to incorporate historical data into model parameters. Historical sequence-based sampling (Yao et al. 2022) selects samples based on the trends and variances of historical uncertainty trends, while TiDAL (Kye et al. 2023) predicts dynamic trends to measure uncertainty.

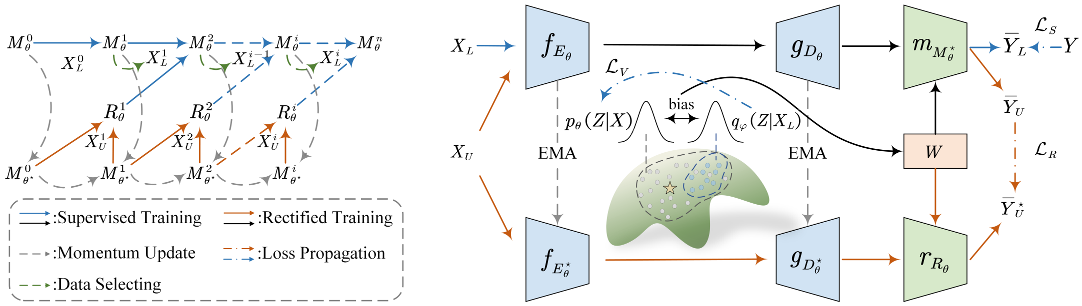

Figure 2: Overall process (left) and framework (middle) of VaGeRy. After a round of supervised training, a rectified training round is performed, during which the bias from the previous round is "saved" for rectification and subsequently "received" into the next round of supervised training. The bias variable serves as a measure of bias, quantifying the discrepancy between the posterior distribution of the current labeled samples and an approximate overall posterior distribution. After passing through the rectification memorization module, it informs the training of the rectifier; upon traversing the rectification receptance module, it guides the rectification of the supervised training process.

Consistency Constraints Semi-supervised learning (SSL) and active learning share similar goals and can complement each other. Bias phenomenon exists as well in SSL (Chen et al. 2022), characterized as the "Matthew Effect". To address bias in few-shot learning, many semi-supervised active learning methods integrate SSL and active learning, using active learning to collect hard samples from unlabeled data and SSL to harness consistency constraints for model training (Gao et al. 2020; Guo et al. 2021; Huang et al. 2024, 2021; Hwang et al. 2023). Some studies also use pseudo-labeling methods to reduce bias from hard sample collection in active learning (Elezi et al. 2022).

Latent Space Representation, widely used in Deep Learning (Goodfellow et al. 2014; Ho, Jain, and Abbeel 2020; Kingma and Welling 2014), embeds both labeled and unlabeled data into a latent space, primarily using VAEs (Kingma and Welling 2014) in adversarial active learning methods (Deng et al. 2018; Geng, Liu, and Qin 2023; Guo et al. 2021; Liu et al. 2020; Sinha, Ebrahimi, and Darrell 2019). These methods focus on using latent representations for selecting hard samples without incorporating them into model training. Despite task-aware approaches like TA-VAAL (Kim et al. 2021) and SRAAL (Zhang et al. 2020), the potential of latent variable biases in enhancing model training is overlooked.

Methodology

This paper proposes an active learning framework based on Variational Gradient Rectifier (VaGeRy). We first introduce the process of rectification and the working principles of the overall framework; Then we discuss the core component of the rectification framework: the latent space representation based variational bias quantification; And finally, we present the objective functions for training the rectifier: Uncertainty Consistency Regularization and Fluctuation Restriction.

Overview of Historical Ensembled Rectifier

As shown in Fig. 2, in each round of rectification iteration, the process begins with training the Variational Gradient Rectifier R_{φ}^{i-1} on the unlabeled data X_U based on the existing model and its historical versions $\{M_\theta^j\}_{j=0}^{i-1}$. This stage is referred to as *the rectification memorization phase* (indicated by the orange arrow in Fig. 2). Next, the model M_θ^{i-1} is trained on X_L^{i-1} based on R_{φ}^{i-1} to obtain M_θ^i, which rectifies for biases in historical models from previous active learning iterations while embarking on a new round of iteration. This stage is called *the rectification reception phase* (indicated by the blue arrow in Fig. 2). Finally, based on M_θ^i and a specific uncertainty measurement method, the most informative unlabeled data is selected from the unlabeled data pool and sent to oracle for labeling, updating the labeled data pool X_L^i. This stage is referred to as *the data selecting and labeling phase* (indicated by the green arrow in Fig. 2). Let $M_{\theta^\star}^{i-1}$ denotes the historical baseline model $\{M_\theta^j\}_{j=0}^{i-1}$, with the gray arrow representing its historical ensembling. Fig. 2 provides a detailed structure of VaGeRy, where M_θ is disentangled into encoder f_{E_θ} and decoder g_{D_θ}. Similarly, $M_{\theta^\star}$ is disentangled into $f_{E_\theta^\star}$ and $g_{D_\theta^\star}$. The Rectifier consists of three parts: the Memory module m_{R_θ}, the bias Variable W, and the Receptance module r_{R_θ}. The behaviors of the three modules are detailed differently three stages below:

The Rectification Memorization Phase: The bias generated during the training of M_θ^{i-1} on X_L^{i-2} is "saved" in the Memory module after being measured on X_U against $\{M_\theta^j\}_{j=0}^{i-1}$ and updated via stochastic gradient descent. Similar to the approach in (Baevski et al. 2022; Grill et al. 2020; He et al. 2020; Huang et al. 2021; Tarvainen and Valpola 2017), this paper also employs exponential moving average for momentum updating of the historical model $M_{\theta^\star}$: $\theta_i^\star = m \times \theta_{i-1}^\star + (1-m) \times \theta_i$ where m is the EMA decay rate. Subsequently, m_{R_θ} is updated using the rectified loss $\mathcal{L}_R$ (introduced in Eq. (13)). Specifically, the objective is optimized as follows:

$$\min_{R_\theta} \mathbb{E}_{z_U \sim p_{x_U}} \mathbb{E}_{z_U^\star \sim p_{x_U}^\star} \Big[\mathcal{L}_R \big[m_{R_\theta} \left(g_{D_\theta^\star} \left(z_U^\star \right), w_U^\star \right), \\ r_{R_\theta} \left(g_{D_\theta} \left(z_U \right), w_U \right) \big] \Big] \tag{1}$$

where $z_U \sim p_{x_U}$ represents $z_U = f_\theta \left(x_U \right), x_U \sim X_U$, and $z_U^\star \sim p_{x_U}^\star$ represents $z_U^\star = f_{\theta^\star}(x_U), x_U \sim X_U$. w_U and $w_U^\star$ serve as references for rectification, constrained by a variational loss $\mathcal{L}_V$, specifically optimizing the following objective:

$$\min_{R_\theta} \mathbb{E}_{z_U \sim p_{x_U}} \left[\mathcal{L}_V \left(z_u, w_u \right) \right] \tag{2}$$

which will be detailed in Eq. (10).

The Rectification Reception Phase: The historical rectification information "saved" in the Memory module is "received" into the new round of training for M_θ^{i-1} based on X_L^{i-1}. the supervised task loss $\mathcal{L}_S$ is used to update M_θ^{i-1} and r_{R_θ}, optimizing the following objective:

$$\min_{M_\theta^i, R_\theta} \mathbb{E}_{z_L \sim p_{x_L}} \Big[\mathcal{L}_S \big(r_{R_\theta} \left(g_{D_\theta} \left(z_L \right), w_L \right), y_L \big) \Big] \tag{3}$$

where $z_L \sim p_{x_L}$ represents $z_L = f_\theta(x_L), x_L \sim X_L^{i-1}$. w_L also serves as a reference for rectification. For a classification task, $\mathcal{L}_S$ is the cross-entropy loss.

The Data Selecting and Labeling Phase: It is fundamentally consistent with existing active learning methods, with the distinction that our method applies a rectification to the predictions. The probability prediction of M_θ^i for an unlabeled sample x_U is given by:

$$\hat{y}_U^{(i)} = \Big[p^{(i)}(1 \mid x_U), p^{(i)}(2 \mid x_U) \cdots, p^{(i)}(C \mid x_U) \Big] \tag{4}$$

where C is the number of classes in the dataset. In existing active learning methods, the uncertainty of a sample x_U under M_θ^i is measured using $\hat{y}_U^{(i)}$, such as the Entropy method H (Shannon 1948). The samples with the highest uncertainty are sorted and the top k samples are selected for labeling by the oracle:

$$X_L^i = X_L^{i-1} \cup \Big\{ \text{argTopk} H(\hat{y}_U) \Big\} \tag{5}$$

In VaGeRy, we apply the Rectifier for rectification, obtaining the rectified probability prediction M_θ^i for an unlabeled sample x_U: $r_{R_\theta}\left(p^{(i)}(1 \mid x_U), w_U\right)$. Based on $\bar{y}_U^{(i)}$, we measure the uncertainty of a sample x_U under M_θ^i and obtain the rectified selection:

$$\begin{aligned} X_L^i &= X_L^{i-1} \cup \Big\{ \text{argTopk} H\left(\hat{y}_U^{(i)}\right) \Big\} \\ &= X_L^{i-1} \cup \Big\{ \text{argTopk} - \sum_{c=1}^{C} \Big[r_{R_\theta}\left(p^{(i)}(c \mid x_U), w_U \right) \\ &\quad \times \log r_{R_\theta}\left(p^{(i)}(c \mid x_U), w_U \right) \Big] \Big\} \end{aligned} \tag{6}$$

Our method is a combination of active learning and SSL, but unlike (Elezi et al. 2022; Huang et al. 2024) , VaGeRy operates through an alternating iteration, delayed rectification process. Furthermore, most SSL methods indiscriminately apply consistency regularization signals obtained X_U and $\mathcal{L}_S$ obtained from X_L in an equivalent manner to gradient updates of parameters. However, there inevitably exists a "local-global" or "bias-generalization" difference between X_L and X_U, our paper introduces the Memory module and Receptance module to allow the model to asynchronously and adaptively save and receive signals for bias rectification from X_U.

Variational Bias Quantification

The task of obtaining the posterior distribution of the data domain's generality is intricate, and the representation of latent spaces via variational methods has been demonstrated to yield significant outcomes (Kingma and Welling 2014). The application of latent space representations in active learning, through the lens of variational adversarial methods (Kim et al. 2021; Sinha, Ebrahimi, and Darrell 2019; Zhang et al. 2020), marks a novel approach. Yet, their primary focus remains on optimizing the structure of the selected samples, overlooking the intrinsic value of latent spaces in training processes. This paper endeavors to construct a reference for bias rectification based on latent spaces, specifically, the bias variables W mentioned in the aforementioned framework.

Our objective is to approximate the deviation of the current labeled data pool's distribution from the overall distribution within the data domain, leveraging the extensive pool of unlabeled data that remains untapped. Denoting Z as the latent variable, we aim to optimize the following Variational Lower Bound:

$$\begin{aligned} &\mathbb{E}_{x_L, y_L \sim X_L, Y_L} \mathbb{E}_{q_\varphi(z_L \mid x_L)} \big[\log p_\theta(y_L \mid z_L) \big] \\ &\qquad - \beta KL\big[q_\varphi(z_L \mid x_L) \parallel p_\theta(\tilde{z}) \big] \\ &+ \mathbb{E}_{x_U \sim X_U} \mathbb{E}_{q_\varphi(z_U \mid x_U)} \Big[KL\big[p_\theta^\star(y_U^\star \mid z_U^\star) \parallel p_\theta(y_U \mid z_U) \big] \Big] \\ &\qquad - \beta KL\big[q_\varphi(z_U \mid x_U) \parallel p_\theta(\tilde{z}) \big] \end{aligned} \tag{7}$$

where p_z represents the probability distribution of z, and KL denotes the Kullback–Leibler (KL) divergence. Upon determining p_z, we define $w_U^\star \triangleq z - z_U^\star$, $w_U \triangleq z - z_U$ in Eq. (1), and $w_L \triangleq z - z_L$ in Eq. (3). We further delineate the calculation of the rectified prediction distribution during the rectification memorization phase as:

$$\begin{aligned} \bar{y}_U^\star &= m_{R_\theta}\left(g_{D_\theta^\star}(z_U^\star), w_U^\star \right) = \text{MLP}(w_U^\star) + \hat{y}_U^\star \\ \bar{y}_U &= r_{R_\theta}\left(g_{D_\theta}(z_U), w_U \right) = \text{MLP}(w_U) + \hat{y}_U \end{aligned} \tag{8}$$

and during the reception rectification phase as:

$$\bar{y}_L = r_{R_\theta}\left(g_{D_\theta}(z_L), w_L \right) = \text{MLP}(w_L) + \hat{y}_L \tag{9}$$

Disentangling the supervised loss $\mathcal{L}_S$ and the consistency loss (Rectified Loss $\mathcal{L}_R$, which is modified in the following section) from Variational Lower Bound, we ultimately arrive at the variational loss $\mathcal{L}_V$ expressed as:

$$\mathcal{L}_V = \mathbb{E}_{x \sim X} \mathbb{E}_{q_\varphi(z \mid x)} KL\big[q_\varphi(z \mid x) \parallel p_\theta(\tilde{z}) \big] \tag{10}$$

Rectification Memorization Objectives

In this section, we provide a detailed discussion on the rectified loss $\mathcal{L}_R$ mentioned in Eq. (1). Most existing methods apply consistency regularization at the distribution level (Gao et al. 2020; Hekimoglu et al. 2023; Roth and Small 2006), neglecting more fine-grained level consistency measures. The characteristics of training dynamics (Kye et al. 2023; Swayamdipta et al. 2020; Yao et al. 2022) are typically utilized for ranking the informational content of unlabeled samples in active learning. Training Dynamics can also be leveraged to alleviate the bias issues associated with active learning. The designed rectified loss $\mathcal{L}_R$ in this

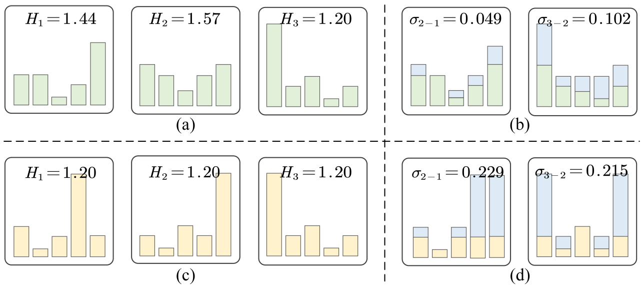

Figure 3: The training dynamics delineate two contradictions often overlooked in conventional active learning. (a) and (b) respectively present the uncertainty assessment of a certain sample over three iterations and the standard deviation between these assessments. While the uncertainty varies across each iteration, it ultimately converges to 1.2, with a relatively small standard deviation. Similarly, (c) and (d) offer analogous assessments, yet the uncertainty remains constant at 1.2 across three iterations, albeit with a larger standard deviation. (Chen et al. 2022) has also noticed similar contradictions.

paper comprises two parts: **Uncertainty Consistency Regulation (UCR)** and **Fluctuation Restriction (FR)**.

UCR measures consistency at the uncertainty level, serving as a holistic metric. As depicted in Fig. 3(a), for conventional uncertainty measures that do not employ Temporal Difference principles (e.g., entropy), a sample exhibiting high predictive probability for a particular class and low overall instability during the third measurement round may appear to be a straightforward, easily learnable instance with limited utility (Swayamdipta et al. 2020). However, upon scrutiny of its historical dynamics, substantial inconsistencies are revealed. Conventional semi-supervised methods typically impose distribution-level consistency constraints to enhance model generalization on unlabeled data. In contrast, UCR aims to ensure consistency in uncertainty estimates between M_θ^i and $M_\theta^\star$, thereby mitigating bias. We optimize the following objective:

$$\mathcal{L}_{UCR} = \min_{R_\theta} \mathbb{E}\left[\max\left(\frac{H\left(\bar{y}_U^\star\right)}{H\left(\bar{y}_U\right)+\epsilon_u}, 1\right)\right] \tag{11}$$

where ϵ_u is the inconsistency relaxation coefficient, a small scalar. It's important to note that, as training progresses, the overall uncertainty of X_U on M_θ^i tends to increase and converge. Hence, M_θ^i and $M_\theta^\star$ should not strictly align in their uncertainty measures of X_U. ϵ_u serves as a hyperparameter to control convergence. Eq. (10) only generates backpropagated gradients when $H\left(\bar{y}_U\right) < H(\bar{y}_U^\star) - \epsilon_u$, since $H\left(\bar{y}_U\right) \geq H(\bar{y}_U^\star)$ represents the "convergence of the model", termed "benign inconsistency." This expectation is relaxed to $H\left(\bar{y}_U\right) \geq H(\bar{y}_U^\star) - \epsilon_u$, while $H\left(\bar{y}_U\right) < H(\bar{y}_U^\star) - \epsilon_u$ is considered as "malignant inconsistency", necessitating rectification.

FR measures consistency at the class level, representing a local metric. Even if certain samples exhibit consistency in uncertainty measurement, their historical measures might display varying statistical properties or fluctuations. As illustrated in Fig. 3(c), across three historical iterations, an unlabeled sample is predicted with consistently low instability but different outcomes. For methods that do not consider Training Dynamics and even for the perspective of UCR, this sample would not be classified as hard across these rounds. However, as shown in Fig. 3(d), there is significant fluctuation (standard deviation) between trainings. (Kye et al. 2023; Yao et al. 2022) utilizes such fluctuations to select hard samples. we aim to rectify training bias by limiting fluctuations at the class level. Specifically, we optimize the following objective:

$$\mathcal{L}_{FR} = \min_{R_\theta} \mathbb{E}\left[\max\left(\sqrt{\mathrm{Var}\left(\bar{y} - \bar{y}_U^\star\right)}, \epsilon_f\right)\right] \tag{12}$$

where ϵ_f is the fluctuation relaxation coefficient, a small scalar hyperparameter similar to ϵ_u. We seek to limit the fluctuation between M_θ^i and $M_\theta^\star$ in their predictions of X_U: when fluctuations occur, even to the point of inconsistency between $\arg\max_{c=1,2\cdots C} m_{R_\theta} \cdot \left(p\left(c \mid x_U\right), w_U^\star\right)$ and $\arg\max_{c=1,2\cdots C} r_{R_\theta} \cdot \left(p\left(c \mid x_U\right), w_U\right)$, the fluctuation variance $\mathrm{Var}(\bar{y} - \bar{y}_U^\star)$ presents higher values. Corrective gradients are backpropagated when this exceeds a threshold ϵ_f. Finally, the rectified loss $\mathcal{L}_R$ is defined as:

$$\mathcal{L}_R = \lambda \mathcal{L}_{UCR} + \delta \mathcal{L}_{FR} \tag{13}$$

where λ and δ are hyperparameters.

Experiments

In this section, we subject VaGeRy to empirical scrutiny. We first provide a detailed account of the datasets employed, baseline methodologies against which comparisons are drawn, and the experimental configurations established. Then we undertake bias rectification assessments on six classical uncertainty-aware techniques and SOTA active learning methodologies across six benchmark datasets for text classification. We also compare VaGeRy with SOTAs on image classification task. Next, we conduct comparative evaluations of SOTA performance under extreme bias conditions, particularly in the context of out-of-distribution (OOD) generalization tasks over three additional benchmark datasets for text classification. Lastly, we deliberate upon the inherent limitations of our proposed methodology. We perform extensive ablation study to elucidate the influence of individual modules within VaGeRy in the Appendix.

Experimental Setup

Datasets. We conduct bias rectification validations on six text datasets: SST-2 (Socher et al. 2013), AGNews (Zhang, Zhao, and LeCun 2015), PubMed (Dernoncourt and Lee 2017), TREC (Li and Roth 2002), SST-5 (Socher et al. 2013) and Newsgroup (Lang 1995). We also conduct experiments on two image datasets: Cifar-10 and Cifar-100 (Krizhevsky and Hinton 2009). For OOD generalization tasks, we employ the same datasets as in (Deng et al. 2023): SA (Kaushik, Hovy, and Lipton 2020), NLI (Kaushik, Hovy, and Lipton 2020), and ANLI (Houlsby et al. 2011). However, different from (Deng et al. 2023), we concatenate the unlabelled

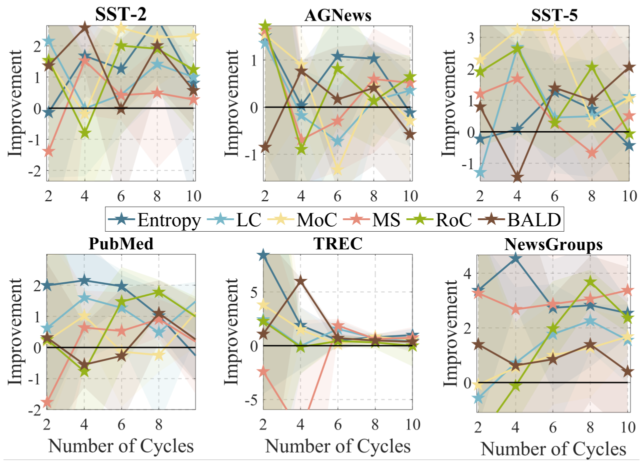

Figure 4: The relative improvement of six uncertainty-aware methods rectified by VaGeRy on the SST-2, AGNews, SST-5, PubMed, TREC, and Newsgroup datasets. **Detailed comparisons are provided in the Appendix**.

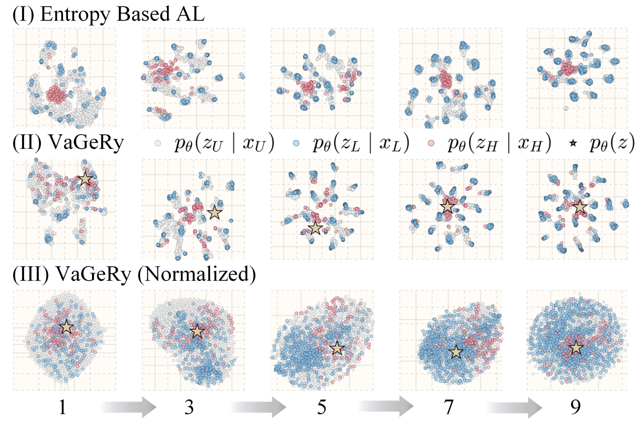

Figure 5: A T-SNE visualization of the iterative active learning process based on entropy, along with its rectification by VaGeRy, exemplified on the Newsgroup dataset.

portions of the Twitter (Rosenthal, Farra, and Nakov 2017) dataset with those of SA to form the unlabelled data used for training the rectifier (with each new subset of the hard sample selected per round still originating from the unlabelled portion of SA). We merge the unlabelled portions of ANLI and NLI with each other's respective unlabelled sets to constitute the unlabelled data for rectifier training.

Baselines. We choose classical uncertainty-aware methods: **MoC** (Cao et al. 2019; Pleiss et al. 2020; Roth and Small 2006), **LC** (Culotta and McCallum 2005), **RoC** (Mosqueira-Rey et al. 2023), **Entropy** (Shannon 1948; Wang et al. 2023; Wu, Chen, and Huang 2022); diversity-aware methods: **Core-set** (Sener and Savarese 2018), **Core-GCN** (Caramalau, Bhattarai, and Kim 2021) alongside the **BALD** (Houlsby et al. 2011), **Learning Loss** (Yoo and Kweon 2019) and the SOTA methods **NoiseStability** (Li et al. 2024), **TiDAL** (Kye et al. 2023) as baselines, integrating our approach with these to mitigate their inherent biases. For the OOD generalization task, we compare our VaGeRy method with the **CounterAL** (Deng et al. 2023) method.

Evaluation details. To ensure a fair comparison, we fix five random seeds. After reproducing the baseline methods, we integrate them into our framework for rectification. We compute the mean accuracy across five trials with 95% confidence intervals. Each experiment consists of ten iterations (cycles); however, due to consistent initial labeled samples, we discard the first round and, for brevity, report accuracies at checkpoints 2, 4, 6, 8, and 10. For images classification task, we report checkpoints from 1 to 7. Implementation details are provided in the Appendix.

Results on Uncertainty-Aware Methods

Improvement on text classification task. Figure 4 showcases the improvements achieved by augmenting six classical uncertainty-aware methods with our VaGeRy approach across six datasets. Overall, our method yields enhancements in almost all 36 configurations, encompassing 80% checkpoint uplifts. We compare the SOTA method NoiseStability (Li et al. 2024) with our approach on two datasets, as illustrated in Figure 6, where our method demonstrates significant improvement in 80% of the iterations (9 checkpoints in 10 for SST-5 and 7 checkpoints in 10 for TREC). These observations suggest that the enhancements our method brings to different baselines are contingent upon the strength of the baselines themselves. More intuitive results can be found in Table 1 in the Appendix.

Improvement on image classification task. We also compare our method with the SOTAs on the Cifar-10 and Cifar-100 datasets, as shown in Figure 7. Our Entropy based version of VaGeRy outperforms the diversity-aware based CoreSet (Sener and Savarese 2018), Core-GCN (Caramalau, Bhattarai, and Kim 2021) and predicting-auxiliary based Learning Loss (Yoo and Kweon 2019) methods and even comparable to TiDAL (Kye et al. 2023). We assume that diversity-aware methods comparatively crude since they strive for unbiasedness at the data collection level instead of the post-prior distribution level.

Visualizations and analysis. We further employ t-SNE mapping to visually demonstrate the effect of our VaGeRy rectification on an entropy-based uncertainty-aware active learning method applied to the Newsgroups dataset, as depicted in Figure. 5. In the entropy-based approach, hard samples selected during each iteration exhibit a pronounced tendency to cluster, reflecting the model's preference for what it deems as "most informative" examples at that moment. Post-VaGeRy rectification, the degree of clustering among hard samples diminishes, with these instances now predominantly scattered across ambiguous regions between clusters rather than aggregating as a whole. Additionally, we visualize the representations in the normalized latent space (i.e., $q_\varphi(z \mid x)$), revealing a similarly widespread distribution of the selected hard samples. Despite exhibiting diminished discriminative information after normalization, the corresponding accuracy indeed improves, as evidenced in Table 1.

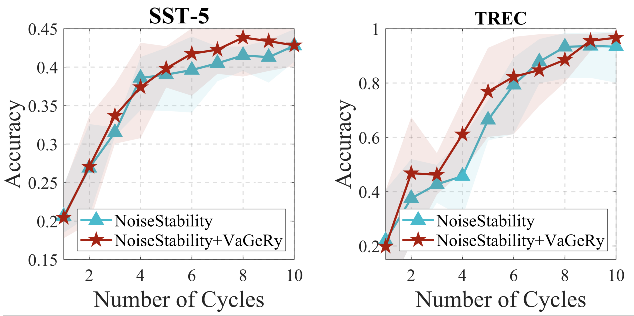

Figure 6: The performance enhancement resulting from VaGeRy's rectification of NoiseStability on the SST-5 and TREC datasets.

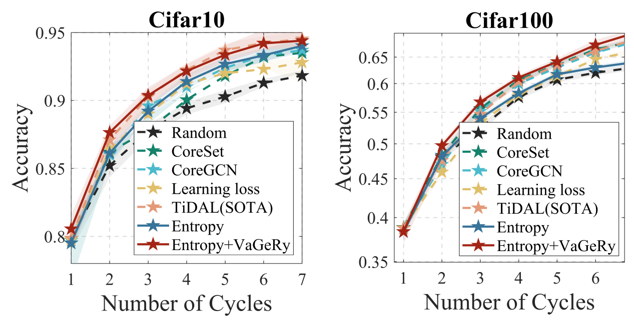

Figure 7: The performance comparison between Entropy based VaGeRy and current SOTAs on Cifar-10 and Cifar-100 datasets.

Out of Distribution Generalization

In the extreme case of bias, OOD generalization tasks, we contrast the performance of SOTA methods CounterAL(Deng et al. 2023). We posit that when statistical bias manifests as a disparity between independent and identically distributed (IID) and OOD data, models may suffer from poor OOD generalization due to preference for the IID domain. VaGeRy enhances model performance across both domains by its comprehensive understanding of both IID and OOD data. **Improvement.** Figure 8 presents the comparison of CounterAL and its rectified version under VaGeRy on three datasets. For the SA and ANLI datasets, VaGeRy models outperform CounterAL, whereas for the NLI dataset, improvements are observed in the early and late rounds, but intermediate rounds show suboptimal performance. We speculate that this could be attributed to an inadequately controlled ratio of unlabeled data used in the process.

Visualizations and analysis. On the ANLI dataset, we employ t-SNE mapping to visually illustrate the model's behavior throughout the active learning iterative process. Figure 9 reveals that upon the integration of VaGeRy into CounterAL, a marked enhancement in the clustering characteristics within the latent space is observed as the cycles progress, with the labeled samples attaining greater representativeness. This indicates that VaGeRy amplifies the agglomeration of akin samples in the latent space, thereby bolstering the algorithm's precision and generalization capabilities in the context of unlabeled data. Evident from the visual representation is the notable refinement in cluster definition and delineation, especially between the 7th and 9th cycles, which suggests that VaGeRy incrementally refines the model's internal representations, thus escalating the distinctiveness of the samples.

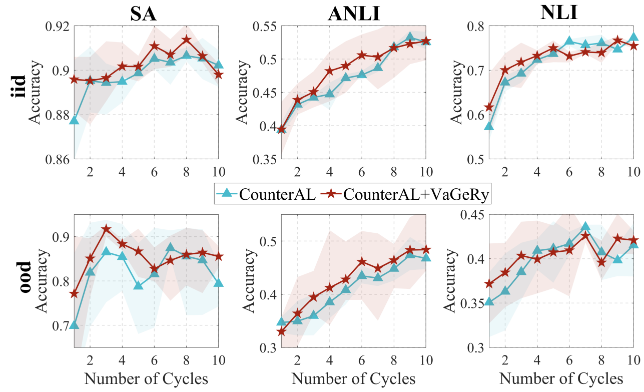

Figure 8: A comparison of iid and ood performance based on CounterAL, with and without VaGeRy rectification, as observed on the SA, ANLI, and NLI datasets.

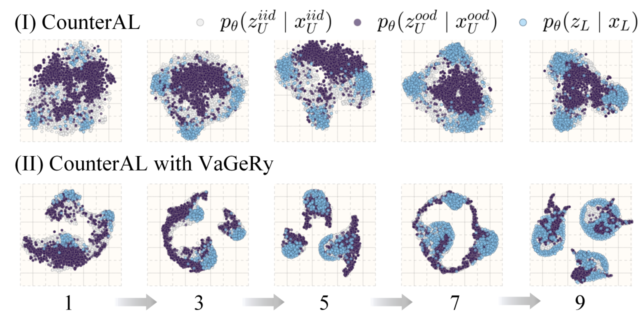

Figure 9: A comparison of T-sne visualization of iterative latent state optimization between (I) CounterAL algorithm across different data distributions and (II) CounterAL rectified by VaGeRy, demonstrating how different iteration cycles affect the latent representation of data, leading to a clearer separation of distinct data distributions.

Conclusion

This paper revisits the inherent bias issue in active learning, analyzing its suppressive effect on active learning performance. In response, we propose a framework named VaGeRy, which dynamically adaptively rectifies bias to unlock the full potential of active learning. VaGeRy consists of two main components: a bias measurement based on variational methods, and an Uncertainty Consistency Regularization and Fluctuation Restriction that integrates training dynamics. Focusing on ten text classification datasets and two image classification datasets, extensive experiments demonstrats the efficacy of VaGeRy in mitigating bias within active learning scenarios.

Dataset	Model	Cycles					Dataset	Model	Cycles				
		2-th	4-th	6-th	8-th	10-th			2-th	4-th	6-th	8-th	10-th
SST-2	Entropy	83.78	84.40	86.52	84.54	85.77	AGNews	Entropy	80.10	87.77	87.43	87.62	88.76
	+VaGeRy	83.64	86.08	87.78	88.30	86.57		+VaGeRy	81.61	87.81	88.51	88.65	88.65
	Improvement	**-0.13**	**1.68**	**1.26**	**3.76**	**0.80**		**Improvement**	**1.52**	**0.03**	**1.08**	**1.03**	**-0.12**
	LC	81.50	85.69	85.99	85.70	86.59		LC	79.84	87.88	88.20	88.28	87.64
	+VaGeRy	83.65	85.68	86.42	88.08	87.59		+VaGeRy	81.19	87.70	87.47	88.42	88.00
	Improvement	**2.15**	**-0.01**	**0.43**	**2.38**	**1.00**		**Improvement**	**1.35**	**-0.18**	**-0.73**	**0.14**	**0.36**
	MoC	82.17	85.56	84.73	85.50	85.31		MoC	80.11	87.00	88.80	88.15	88.27
	+VaGeRy	83.65	85.38	87.31	87.76	87.63		+VaGeRy	81.65	87.90	87.47	88.62	87.98
	Improvement	**1.48**	**-0.18**	**2.58**	**2.27**	**2.32**		**Improvement**	**1.54**	**0.9**	**-1.33**	**0.47**	**-0.29**
	MS	82.42	85.05	86.26	86.45	87.56		MS	76.43	88.17	88.35	87.46	88.21
	+VaGeRy	81.03	86.58	86.67	86.94	87.83		+VaGeRy	78.06	87.47	85.84	88.05	88.72
	Improvement	**-1.38**	**1.52**	**0.42**	**0.49**	**0.28**		**Improvement**	**1.63**	**-0.70**	**-2.51**	**0.60**	**0.52**
	RoC	82.17	85.56	84.73	85.49	85.31		RoC	81.21	87.95	87.46	88.01	87.33
	+VaGeRy	83.71	84.76	86.74	87.40	86.55		+VaGeRy	82.93	87.06	88.28	88.14	87.98
	Improvement	**1.54**	**-0.80**	**2.01**	**1.91**	**1.24**		**Improvement**	**1.72**	**-0.90**	**0.82**	**0.13**	**0.65**
	BALD	81.70	83.92	85.97	84.91	87.29		BALD	68.98	86.12	87.92	88.21	88.53
	+VaGeRy	83.06	86.50	85.94	88.20	87.85		+VaGeRy	68.13	86.89	88.09	88.61	87.96
	Improvement	**1.36**	**2.58**	**-0.03**	**3.30**	**0.57**		**Improvement**	**-0.85**	**0.77**	**0.17**	**0.41**	**-0.58**
SST-5	Entropy	38.56	43.46	45.86	45.23	47.72	PubMed	Entropy	69.64	77.62	78.97	79.16	79.57
	+VaGeRy	38.34	43.54	47.13	47.32	47.29		+VaGeRy	71.64	79.77	79.53	80.18	79.29
	Improvement	**-0.23**	**0.08**	**1.27**	**2.09**	**-0.44**		**Improvement**	**2.00**	**2.16**	**0.56**	**1.02**	**-0.28**
	LC	40.07	43.99	46.78	47.31	47.65		LC	70.45	77.34	79.36	79.49	79.48
	+VaGeRy	38.78	46.67	47.23	47.80	48.77		+VaGeRy	71.08	78.94	80.63	80.98	80.98
	Improvement	**-1.29**	**2.68**	**0.45**	**0.50**	**1.12**		**Improvement**	**0.63**	**1.60**	**1.27**	**0.48**	**1.50**
	MoC	39.62	43.90	45.18	47.44	47.51		MoC	70.83	78.99	79.83	80.78	79.35
	+VaGeRy	41.90	45.33	48.41	47.75	48.58		+VaGeRy	71.04	80.01	79.70	80.54	80.50
	Improvement	**2.29**	**1.43**	**3.23**	**0.31**	**1.06**		**Improvement**	**0.21**	**1.02**	**-0.13**	**-0.25**	**1.15**
	MS	37.57	40.52	41.57	43.62	44.83		MS	70.26	78.93	79.59	79.79	80.36
	+VaGeRy	38.78	42.21	41.84	42.94	45.34		+VaGeRy	68.50	79.57	80.12	80.70	80.54
	Improvement	**1.21**	**1.68**	**0.27**	**-0.68**	**0.51**		**Improvement**	**-1.76**	**0.64**	**0.53**	**0.91**	**0.18**
	RoC	39.83	44.05	47.14	46.61	48.23		RoC	70.53	79.57	78.42	79.01	80.09
	+VaGeRy	41.73	46.69	47.42	48.67	48.16		+VaGeRy	70.79	78.80	79.90	80.80	81.08
	Improvement	**1.91**	**2.64**	**0.28**	**2.06**	**-0.06**		**Improvement**	**0.25**	**-0.77**	**1.47**	**1.78**	**0.98**
	BALD	37.92	39.97	38.73	41.67	42.10		BALD	70.26	78.93	79.59	79.79	80.36
	+VaGeRy	38.71	38.53	40.98	42.66	44.15		+VaGeRy	70.57	78.38	79.32	80.89	80.61
	Improvement	**0.79**	**-1.44**	**2.25**	**0.99**	**2.06**		**Improvement**	**0.32**	**-0.55**	**-0.27**	**1.10**	**0.24**
TREC	Entropy	58.88	94.26	97.20	98.16	97.65	Newsgroups	Entropy	24.78	49.37	56.65	59.01	60.80
	+VaGeRy	67.30	96.17	97.70	98.95	98.63		+VaGeRy	28.15	53.91	59.38	61.83	63.34
	Improvement	**8.42**	**1.90**	**0.50**	**0.79**	**0.98**		**Improvement**	**3.37**	**4.54**	**2.73**	**2.83**	**2.54**
	LC	52.67	96.66	96.73	98.44	98.24		LC	30.23	52.33	58.46	62.33	63.01
	+VaGeRy	55.10	96.63	98.29	99.13	98.70		+VaGeRy	29.66	53.05	60.23	64.59	64.56
	Improvement	**2.42**	**-0.03**	**1.56**	**0.69**	**0.46**		**Improvement**	**-0.57**	**0.72**	**1.77**	**2.26**	**1.55**
	MoC	60.08	94.98	98.39	97.73	98.10		MoC	32.71	54.24	58.51	61.74	62.43
	+VaGeRy	63.87	96.47	98.53	98.54	98.68		+VaGeRy	32.62	54.79	59.51	63.00	64.10
	Improvement	**3.80**	**1.49**	**0.14**	**0.81**	**0.58**		**Improvement**	**-0.08**	**0.55**	**1.00**	**1.27**	**1.68**
	MS	56.37	95.25	95.95	97.39	97.69		MS	21.90	39.30	48.50	53.98	56.97
	+VaGeRy	53.96	87.22	97.82	98.00	98.43		+VaGeRy	25.16	41.97	51.37	57.04	60.34
	Improvement	**-2.41**	**-8.03**	**1.87**	**0.61**	**0.74**		**Improvement**	**3.27**	**2.68**	**2.87**	**3.06**	**3.37**
	RoC	62.15	96.91	98.44	98.16	98.01		RoC	33.92	54.59	59.88	61.23	62.98
	+VaGeRy	64.44	96.79	98.87	98.46	98.00		+VaGeRy	31.86	54.47	61.85	64.91	65.34
	Improvement	**2.28**	**-0.12**	**0.43**	**0.30**	**-0.01**		**Improvement**	**-2.05**	**-0.12**	**1.97**	**3.68**	**2.36**
	BALD	57.93	91.45	97.12	97.55	97.79		BALD	29.87	48.62	53.56	56.13	59.38
	+VaGeRy	59.02	97.45	97.82	98.03	98.16		+VaGeRy	31.27	49.24	54.41	57.53	59.78
	Improvement	**1.08**	**5.99**	**0.70**	**0.48**	**0.37**		**Improvement**	**1.39**	**0.63**	**0.85**	**1.39**	**0.40**

Table 1: A comparative evaluation of six uncertainty-aware methods and their respective performances under VaGeRy rectification, and associated improvements, conducted across six datasets.

Related Work

Uncertainty-aware methods focus on selecting unlabeled samples with high uncertainty, as they are believed to contain more information for model generalization. The main challenge in this branch is how to assess uncertainty. Classic methods include those based on the margin of confidence (MoC) (Cao et al. 2019; Pleiss et al. 2020; Roth and Small 2006), least confidence (LC) (Culotta and McCallum 2005), entropy-based methods (Shannon 1948; Wang et al. 2023; Wu, Chen, and Huang 2022), and the ratio of confidence (RoC) (Mosqueira-Rey et al. 2023).

Recent advances include predicting-auxiliary or synthesizing methods that learnably measure the contribution of unlabeled data to model convergence, such as loss prediction (Yoo and Kweon 2019), adversarial methods (Deng et al. 2018; Guo et al. 2021; Kim et al. 2021; Sinha, Ebrahimi, and Darrell 2019; Zhang et al. 2020), and dynamic prediction (Kye et al. 2023). Training-Free methods have also shown success (Xie et al. 2023).

Representation-aware approaches aim to represent the entire dataset with a limited number of labeled samples to enhance model generalization on unlabeled data, with CoreSet (Sener and Savarese 2018) being a typical example. Variety-aware methods prioritize selecting a diverse set of samples (Guo 2010).

Variational Lower Bound

Variational Lower Bound in Supervised Learning

The Deep Variational Information Bottleneck (DVIB) (Alemi, Fischer, and Dillon 2017) employs variational techniques to extend the information bottleneck (TISHBY 2000) concept from information theory to deep neural networks, constituting a generalization of the Variational Autoencoder (VAE) (Kingma and Welling 2014). Reference (Mahabadi, Belinkov, and Henderson 2021) applies the Variational Information Bottleneck to low-resource text classification tasks. Specifically, for typical supervised learning contexts, they optimize the following variational lower bound:

$$\mathbb{E}_{x,y\sim X,Y}\mathbb{E}_{q_{\varphi}(z|x)}\left[logp_{\theta}(y\mid z)\right]-\beta KL\left[q_{\varphi}(x\mid z)\parallel p_{\theta}(z)\right] \tag{14}$$

Variational Lower Bound in Active Learning Paradigm

We now consider the variational lower bound within the paradigm of active learning.

$$\begin{aligned}\mathbb{E}_{x_L,y_L\sim X_L,Y_L}\mathbb{E}_{q_{\varphi}(z_L|x_L)}&\left[logp_{\theta}(y_L\mid z_L)\right]\\ &-\beta KL\left[q_{\varphi}(z_L\mid x_L)\parallel p_{\theta}(\hat{z})\right]\end{aligned} \tag{15}$$

we now focus on the empirical evaluation of the variational lower bound on the labeled data, rather than on the overall distribution as in Eq. (14).In active learning scenarios, where labeled data deviates from the global distribution, inherent bias arises. To quantify this bias, we propose measuring the Kullback-Leibler divergence between the parametrized latent variables in the active learning context and those trained under full supervision within the overall distribution.

$$KL\left[p_{\theta}(\hat{z})\parallel p_{\theta}(z)\right]\geq 0 \tag{16}$$

For simplicity, in section 3.2, we introduce bias variable $w_U = z - z_U$ and $w_L = z - Z_L$.

Theorem It can be readily shown that minimizing w is equivalent to minimizing $KL[q_{\varphi}(z|x)\parallel p_{\theta}(\tilde{z})]$ when the bias is minimized.

$$\min\mathbb{E}\big[w\big]\Leftrightarrow\min\mathbb{E}\big[KL\left[q_{\varphi}(z\mid x)\parallel p_{\theta}(\tilde{z})\right]\big] \tag{17}$$

Variational Lower Bound in VaGeRy

In the VaGeRy framework, we jointly minimize the empirical variational lower bound on both labeled and unlabeled datasets, with the latter contributing a consistency regularization constraint.

$$\begin{aligned}&\mathbb{E}_{x_L,y_L\sim X_L,Y_L}\mathbb{E}_{q_{\varphi}(z_L|x_L)}\big[\log p_{\theta}(y_L\mid z_L)\big]\\ &\qquad-\beta KL\big[q_{\varphi}(z_L\mid x_L)\parallel p_{\theta}(\tilde{z})\big]\\ &+\mathbb{E}_{x_U\sim X_U}\mathbb{E}_{q_{\varphi}(z_U|x_U)}\Big[KL\big[p^{\star}_{\theta}(y^{\star}_U\mid z^{\star}_U)\parallel p_{\theta}(y_U\mid z_U)\big]\Big]\\ &\qquad-\beta KL\big[q_{\varphi}(z_U\mid x_U)\parallel p_{\theta}(\tilde{z})\big]\end{aligned} \tag{18}$$

This objective aims to approximate the variational posterior in an unbiased manner.

More Results

Upon application of the VaGeRy method to six distinct datasets and six uncertainty sampling techniques, the comparison of accuracy across various active learning cycles unequivocally demonstrates the method's superior efficacy within the active learning paradigm. As evidenced by Table 1, the VaGeRy method consistently enhances accuracy across a spectrum of active learning cycles, manifesting a progressive amplification of the model's learning efficiency and predictive acumen. Furthermore, the VaGeRy method exhibits a bolstering impact across diverse types of uncertainty estimation techniques, reflecting its stability and adaptability within varying contexts.

Acknowledgments

This research was funded by Research Project BHQ090003000X03, NSFC No.U2441238 and the Science and Technology Innovation Program of Hunan Province No.2023RC1001.

References

Alemi, A. A.; Fischer, I.; and Dillon, J. V. 2017. Deep Variational Information Bottleneck. *ICLR 2017*.

Athiwaratkun, B.; Finzi, M.; Izmailov, P.; and Wilson, A. G. 2019. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In *ICLR 2019*.

Baevski, A.; Hsu, W.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *ICML 2022*, volume 162, 1298–1312.

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS 2019*, 1565–1576.

Caramalau, R.; Bhattarai, B.; and Kim, T. 2021. Sequential Graph Convolutional Network for Active Learning. In *CVPR 2021*, 9583–9592.

Chen, B.; Jiang, J.; Wang, X.; Wan, P.; Wang, J.; and Long, M. 2022. Debiased self-training for semi-supervised learning. *NeurIPS 2022*, 35: 32424–32437.

Culotta, A.; and McCallum, A. 2005. Reducing Labeling Effort for Structured Prediction Tasks. In *AAAI 2005*, 746–751.

Deng, X.; Wang, W.; Feng, F.; Zhang, H.; He, X.; and Liao, Y. 2023. Counterfactual Active Learning for Out-of-Distribution Generalization. In *ACL 2023*, 11362–11377.

Deng, Y.; Chen, K.; Shen, Y.; and Jin, H. 2018. Adversarial Active Learning for Sequences Labeling and Generation. In *IJCAI 2018*, 4012–4018.

Dernoncourt, F.; and Lee, J. Y. 2017. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *IJCNLP 2017*, 308–313.

Elezi, I.; Yu, Z.; Anandkumar, A.; Leal-Taixé, L.; and Álvarez, J. M. 2022. Not All Labels Are Equal: Rationalizing The Labeling Costs for Training Object Detection. In *CVPR 2022*, 14472–14481.

Farquhar, S.; Gal, Y.; and Rainforth, T. 2021. On Statistical Bias In Active Learning: How and When to Fix It. In *ICLR 2021*.

Gao, M.; Zhang, Z.; Yu, G.; Arik, S. Ö.; Davis, L. S.; and Pfister, T. 2020. Consistency-Based Semi-supervised Active Learning: Towards Minimizing Labeling Cost. In *ECCV 2020*, volume 12355, 510–526.

Geng, L.; Liu, N.; and Qin, J. 2023. Multi-Classifier Adversarial Optimization for Active Learning. In *AAAI 2023*, 7687–7695.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NeurIPS 2014*, 2672–2680.

Grill, J.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. Á.; Guo, Z.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *NeurIPS 2020*.

Guo, J.; Shi, H.; Kang, Y.; Kuang, K.; Tang, S.; Jiang, Z.; Sun, C.; Wu, F.; and Zhuang, Y. 2021. Semi-supervised Active Learning for Semi-supervised Models: Exploit Adversarial Examples with Graph-based Virtual Labels. In *ICCV 2021*, 2876–2885.

Guo, Y. 2010. Active Instance Sampling via Matrix Partition. In *NeurIPS 2010*, 802–810.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR 2020*, 9726–9735.

Hekimoglu, A.; Friedrich, P.; Zimmer, W.; Schmidt, M.; Marcos-Ramiro, A.; and Knoll, A. 2023. Multi-Task Consistency for Active Learning. In *ICCV 2023 - Workshops*, 3407–3416.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS 2020*.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Huang, S.; Wang, T.; Xiong, H.; Huan, J.; and Dou, D. 2021. Semi-Supervised Active Learning with Temporal Output Discrepancy. In *ICCV 2021*, 3427–3436.

Huang, S.; Wang, T.; Xiong, H.; Wen, B.; Huan, J.; and Dou, D. 2024. Temporal Output Discrepancy for Loss Estimation-Based Active Learning. *IEEE Trans. Neural Networks Learn. Syst.*, 35(2): 2109–2123.

Hwang, S.; Kim, S.; Kim, Y.; and Kum, D. 2023. Joint Semi-Supervised and Active Learning via 3D Consistency for 3D Object Detection. In *ICRA 2023*, 4819–4825.

Kaushik, D.; Hovy, E. H.; and Lipton, Z. C. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *ICLR 2020*.

Kim, K.; Park, D.; Kim, K. I.; and Chun, S. Y. 2021. Task-Aware Variational Adversarial Active Learning. In *CVPR 2021*, 8166–8175.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR 2014*.

Krishnan, R.; Sinha, A.; Ahuja, N. A.; Subedar, M.; Tickoo, O.; and Iyer, R. 2021. Mitigating Sampling Bias and Improving Robustness in Active Learning. *CoRR*, abs/2109.06321.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images.

Kye, S. M.; Choi, K.; Byun, H.; and Chang, B. 2023. TiDAL: Learning Training Dynamics for Active Learning. In *ICCV 2023*, 22278–22288.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *ICLR 2017*.

Lang, K. 1995. NewsWeeder: Learning to Filter Netnews. In *ICML 1995*, 331–339.

Li, H.; Liu, Y.; Zhang, H.; and Li, B. 2023. Mitigating and Evaluating Static Bias of Action Representations in the Background and the Foreground. In *ICCV 2023*, 19854–19866.

Li, X.; and Roth, D. 2002. Learning Question Classifiers. In *COLING 2002*.

Li, X.; Yang, P.; Gu, Y.; Zhan, X.; Wang, T.; Xu, M.; and Xu, C. 2024. Deep Active Learning with Noise Stability. In *AAAI 2024*, 13655–13663.

Liu, J.; Du, Y.; Xiao, Z.; Snoek, C. G. M.; Sonke, J.; and Gavves, E. 2023. Memory-augmented Variational Adaptation for Online Few-shot Segmentation. In *ICCV 2023 - Workshops*, 3316–3325.

Liu, P.; Wang, L.; Ranjan, R.; He, G.; and Zhao, L. 2022. A Survey on Active Deep Learning: From Model Driven to Data Driven. *ACM Comput. Surv.*, 54(10s): Article 221.

Liu, Y.; Li, Z.; Zhou, C.; Jiang, Y.; Sun, J.; Wang, M.; and He, X. 2020. Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Trans. Knowl. Data Eng.*, 32(8): 1517–1528.

Mahabadi, R. K.; Belinkov, Y.; and Henderson, J. 2021. Variational Information Bottleneck for Effective Low-Resource Fine-Tuning. In *ICLR 2021*.

Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; and Fernández-Leal, Á. 2023. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4): 3005–3054.

Parmar, M.; Mishra, S.; Geva, M.; and Baral, C. 2023. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *EACL 2023*, 1771–1781.

Pleiss, G.; Zhang, T.; Elenberg, E. R.; and Weinberger, K. Q. 2020. Identifying Mislabeled Data using the Area Under the Margin Ranking. In *NeurIPS 2020*.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *ACL 2017*, 502–518.

Roth, D.; and Small, K. 2006. Margin-Based Active Learning for Structured Output Spaces. In *ECML 2006*, volume 4212, 413–424.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *ICLR 2018*.

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

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational Adversarial Active Learning. In *ICCV 2019*, 5971–5980.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP 2013*, 1631–1642.

Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *EMNLP 2020*, 9275–9293.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS 2017*, 1195–1204.

TISHBY, N. 2000. The information bottleneck method. In *ACC, 2000*.

Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Solin, A.; Bengio, Y.; and Lopez-Paz, D. 2022. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145: 90–106.

Wan, T.; Xu, K.; Yu, T.; Wang, X.; Feng, D.; Ding, B.; and Wang, H. 2023. A Survey of Deep Active Learning for Foundation Models. *Intelligent Computing*, 2: 0058.

Wang, Y.; Ilic, V.; Li, J.; Kisacanin, B.; and Pavlovic, V. 2023. ALWOD: Active Learning for Weakly-Supervised Object Detection. In *ICCV 2023*, 6436–6446.

Wu, J.; Chen, J.; and Huang, D. 2022. Entropy-based Active Learning for Object Detection with Progressive Diversity Constraint. In *CVPR 2022*, 9387–9396.

Wu, Z.; Wang, L.; Wang, W.; Xia, Q.; Chen, C.; Hao, A.; and Li, S. 2023. Pixel Is All You Need: Adversarial Trajectory-Ensemble Active Learning for Salient Object Detection. In *AAAI 2023*, 2883–2891.

Xie, Y.; Ding, M.; Tomizuka, M.; and Zhan, W. 2023. Towards Free Data Selection with General-Purpose Models. In *NeurIPS 2023*.

Yao, J.; Dou, Z.; Nie, J.; and Wen, J. 2022. Looking Back on the Past: Active Learning With Historical Evaluation Results. *IEEE Trans. Knowl. Data Eng.*, 34(10): 4921–4932.

Yoo, D.; and Kweon, I. S. 2019. Learning Loss for Active Learning. In *CVPR 2019*, 93–102.

Zhang, B.; Li, L.; Yang, S.; Wang, S.; Zha, Z.; and Huang, Q. 2020. State-Relabeling Adversarial Active Learning. In *CVPR 2020*, 8753–8762.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *NeurIPS 2015*, 28.

Zhang, Z.; Strubell, E.; and Hovy, E. H. 2022. A Survey of Active Learning for Natural Language Processing. In *EMNLP 2022*, 6166–6190.