Leveraging Local Variance for Pseudo-Label Selection in Semi-supervised Learning

Zeping Min[1]*, **Jinfeng Bai**[2], **Chengfei Li**[2]
[1]Peking University, Beijing, China
[2]TAL Education Group, Beijing, China
zpm@pku.edu.cn, baijinfeng1@tal.com, lichengfei@tal.com

Abstract

Semi-supervised learning algorithms that use pseudo-labeling have become increasingly popular for improving model performance by utilizing both labeled and unlabeled data. In this paper, we offer a fresh perspective on the selection of pseudo-labels, inspired by theoretical insights. We suggest that pseudo-labels with a high degree of local variance are more prone to inaccuracies. Based on this premise, we introduce the Local Variance Match (LVM) method, which aims to optimize the selection of pseudo-labels in semi-supervised learning (SSL) tasks. Our methodology is validated through a series of experiments on widely-used image classification datasets, such as CIFAR-10, CIFAR-100, and SVHN, spanning various labeled data quantity scenarios. The empirical findings show that the LVM method substantially outpaces current SSL techniques, achieving state-of-the-art results in many of these scenarios. For instance, we observed an error rate of **5.41%** on CIFAR-10 with a single label for each class, **35.87%** on CIFAR-100 when using four labels per class, and **1.94%** on SVHN with four labels for each class. Notably, the standout error rate of **5.41%** is less than **1%** shy of the performance in a fully-supervised learning environment. In experiments on ImageNet with 100k labeled data, the LVM also reached state-of-the-art outcomes. Additionally, the efficacy of the LVM method is further validated by its stellar performance in speech recognition experiments.

Introduction

The surge in the success of deep learning across diverse domains, such as computer vision (Krizhevsky, Sutskever, and Hinton 2017; Simonyan and Zisserman 2015; Szegedy et al. 2015; He et al. 2017), natural language processing (Mikolov et al. 2013; Raffel et al. 2020; Brown et al. 2020; OpenAI 2023), and speech recognition (Wang et al. 2017; van den Oord et al. 2016; Chan et al. 2015), owes its prosperity to the availability of substantial labeled datasets and powerful computational capabilities. Despite this, the process of acquiring high-quality labeled data frequently proves to be laborious, costly, and time-consuming. Semi-supervised learning (SSL) (Sohn et al. 2020; Chen et al. 2020; Berthelot et al. 2019, 2020) has surfaced as a potent approach to circumvent the need for extensive manual annotation, by exploiting both labeled and unlabeled data to build robust machine learning models. In this context, pseudo-labeling semi-supervised learning algorithms have gained considerable traction.

The essence of pseudo-labeling semi-supervised learning algorithms lies in harnessing the information inherent in the unlabeled data by bestowing pseudo-labels upon them and considering these as additional training instances (Lee et al. 2013; Sohn et al. 2020; Berthelot et al. 2019, 2020). This strategy aids the model in enhancing its generalization capabilities and performance on the task at hand. The efficacy of such approaches hinges on the quality of these pseudo-labels, rendering their selection a paramount aspect of SSL algorithms. Proper pseudo-label selection strategy can bolster the model's ability to generalize to unseen data, while a poor one could precipitate detrimental impacts on model performance.

Prominent semi-supervised learning algorithms that utilize pseudo-labeling, such as MixMatch (Berthelot et al. 2019), FixMatch (Sohn et al. 2020), RemixMatch (Berthelot et al. 2020), and FreeMatch (Wang et al. 2023), have traditionally depended exclusively on the absolute values of posterior probabilities to choose pseudo-labels. These methods interpret these probabilities as a measure of the model's confidence in its predictions, setting a selection threshold accordingly. While this is intuitive and straightforward, it has its limitations.

Specifically, when dealing with sparsely labeled data, the model might predict with high confidence incorrectly, which can adversely affect the performance of these methods. As illustrated in Figure 1, even for a simple binary classification task, relying solely on the model's posterior probabilities for pseudo-label selection can be challenging. This is because many incorrect pseudo-labels might still have high posterior probabilities.

We introduce the Local Variance Match (LVM). This novel method offers a fresh dimension for filtering pseudo-labels. By incorporating the dimension of local variance into our pseudo-label selection, the task becomes considerably more manageable. Removing labels with high local variance helps to reduce errors in pseudo-labels, thereby positively impacting the model's performance (Zhu, Luo, and Liu 2022). Experimental findings showcase that our LVM

*Corresponding Author.

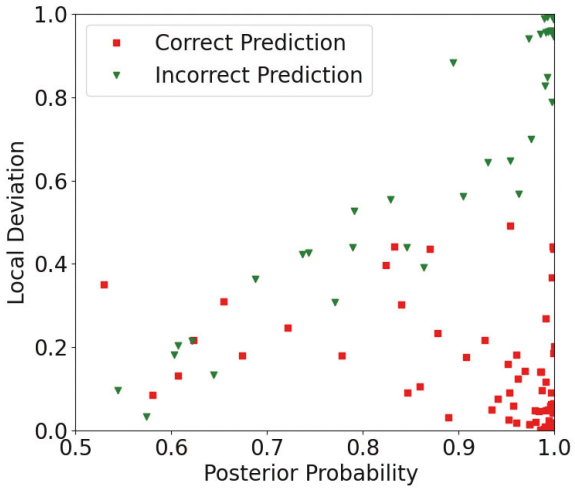

Figure 1: This figure demonstrates the relationship between model predictions, posterior probabilities, and local deviation using a toy experiment. Data points are uniformly sampled within the $[-1.3, 1.3] \times [-1.3, 1.3]$ plane. Points within the unit circle are labeled as 1, while those outside are labeled as 0. We train a 3-layer MLP and use it to predict on test split. The scatter plot illustrates the relationship between model predictions, the predicted posterior probabilities, and the local deviation. Here, 'Local Deviation' for a data point is calculated as the absolute difference between the model's posterior probability of that point and the average predicted probability across its several nearest neighbors for the actual class.

approach outperforms existing SSL methods on popular image classification datasets, such as CIFAR-10, CIFAR-100, and SVHN, achieving state-of-the-art (SOTA) results across various label settings. Notably, LVM achieved an error rate of 5.41% in the most challenging setting of CIFAR-10 with only ten labels (one label per class) – a result that is less than 1% away from the 4.62% error rate observed in a fully-supervised learning setting. As further validation, the LVM method also exhibited superior performance in speech recognition tasks, even with fewer pseudo-labels. These experimental outcomes underscore the efficacy of incorporating local variance into the pseudo-label selection process, leading to improved model performance.

In summary, our paper offers the following significant contributions:

- We present a theoretical analysis to underscore the importance and potential benefits of considering local variance in the pseudo-label selection process for SSL tasks.
- We introduce the Local Variance Match (LVM) method, a novel approach that leverages local variance as a pseudo-label correctness measure to refine pseudo-label selection in SSL tasks.
- We conduct extensive experiments on popular image classification datasets, achieving state-of-the-art results in most settings, and demonstrating better performance in speech recognition tasks with fewer pseudo-labels.

The rest of this paper is structured as follows: Section 2 presents a theoretical analysis underscoring the significance of local variance in pseudo-label selection. In Section 3, we delve into the specifics of the Local Variance Match (LVM) method, explaining its integration of local variance for pseudo-label selection. Section 4 reports the experimental results on image classification and speech recognition datasets, comparing the LVM performance with other SSL methods. Section 5 reviews related work in semi-supervised learning and discusses the existing SSL methods. Finally, Section 6 concludes the paper, highlighting our approach and suggesting avenues for future work.

Motivated Theoretical Analysis

In this section, we delve into a simplified binary classification scenario to shed light on the working mechanics of our data selection strategy, which is rooted in local variance. While our actual model and the associated training process are indeed more sophisticated, this simplified example serves as a conduit for gaining a fundamental understanding of our method.

We consider a binary classification problem where the data distribution is a mixture of two Gaussian distributions. The label Y is equally probable to be positive $(+1)$ or negative (-1). The input x is a scalar and follows the conditional distribution: $x \mid Y = -1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $x \mid Y = +1 \sim \mathcal{N}(\mu_2, \sigma^2)$. Without loss of generality, we assume $\mu_2 > \mu_1$. The optimal classification boundary exists at $(\mu_1 + \mu_2)/2$. We restrict our discussion to the case where $x > (\mu_1 + \mu_2)/2$, where the optimal classifier predicts 1. A similar discussion would hold for $x < (\mu_1 + \mu_2)/2$.

Theorem 1. *Consider a 2-GMM probability model, where $P(Y = i) = \frac{1}{2}$ for $i \in \{-1, +1\}$ and the conditional probability of x follows a normal distribution: $x \mid Y = -1 \sim N(\mu_1, \sigma^2)$ and $x \mid Y = +1 \sim N(\mu_2, \sigma^2)$, where $\mu_1 < \mu_2$. For any point $x > (\mu_1 + \mu_2)/2$, the following properties hold:*

(I) The posterior probability (or the model's confidence level) for a positive class is given by:

$$P(Y = +1 \mid x) = \frac{1}{\exp\left[-\frac{(\mu_2 - \mu_1)\left(x - \frac{\mu_1 + \mu_2}{2}\right)}{\sigma^2}\right] + 1} > P(Y = -1 \mid x) \tag{1}$$

For simplicity, let's denote

$$g(x) \triangleq \exp\left[-\frac{(\mu_2 - \mu_1)\left(x - \frac{\mu_1 + \mu_2}{2}\right)}{\sigma^2}\right] + 1 \tag{2}$$

(II) Differentiating $P(Y = +1 \mid x)$ with respect to x yields:

$$\frac{d}{dx} P(Y = +1 \mid x) = -\frac{1}{g^2(x)} g'(x) > 0 \tag{3}$$

Where

$$g'(x) = -\frac{(\mu_2 - \mu_1)}{\sigma^2} \exp\left[-\frac{(\mu_2 - \mu_1)\left(x - \frac{\mu_1+\mu_2}{2}\right)}{\sigma^2}\right] \tag{4}$$

(III) The local variance of the posterior probability is:

$$\begin{aligned} &Var_{\tilde{x} \overset{uni}{\sim} B_\delta(x)} P(Y = +1 \mid \tilde{x}) \\ &= E_{\tilde{x} \overset{uni}{\sim} B_\delta(x)} P^2(Y = +1 \mid \tilde{x}) \qquad (5) \\ &\quad - \left[E_{\tilde{x} \overset{uni}{\sim} B_\delta(x)} P(Y = +1 \mid x)\right]^2 \\ &= \int_{x-\delta}^{x+\delta} \frac{1}{g^2(\tilde{x})} \frac{1}{2\delta} d\tilde{x} - \left[\int_{x-\delta}^{x+\delta} \frac{1}{g(\tilde{x})} \cdot \frac{1}{2\delta} d\tilde{x}\right]^2 \qquad (6) \end{aligned}$$

(IV) The derivative of $Var_{\tilde{x} \overset{uni}{\sim} B_\delta(x)} P(Y = +1 \mid \tilde{x})$ *with respect to* x *is:*

$$\begin{aligned} &\frac{d}{dx} Var_{\tilde{x} \overset{uni}{\sim} B_\delta(x)} P(Y = +1 \mid \tilde{x}) \\ &= \frac{1}{2\delta}\left(\frac{1}{g(x+\delta)} - \frac{1}{g(x-\delta)}\right) \\ &\quad \times \left(\frac{1}{g(x+\delta)} + \frac{1}{g(x-\delta)} - 2 \cdot \frac{1}{2\delta} \int_{x-\delta}^{x+\delta} \frac{1}{g(\tilde{x})} d\tilde{x}\right) \end{aligned} \tag{7}$$

and when δ *is a small positive quantity, we have*

$$\frac{d}{dx} Var_{\tilde{x} \overset{uni}{\sim} B_\delta(x)} P(Y = +1 \mid \tilde{x}) < 0 \tag{8}$$

From the theorem, several pivotal insights can be distilled:

- The derivative $\frac{d}{dx} P(Y = +1 \mid x)$ is positive. This indicates that an increase in the input variable x corresponds to an increase in the posterior probability $P(Y = +1 \mid x)$ predicted by the model. Consequently, the confidence level of the optimal classifier in its predictions increases as the input progressively diverges from the decision boundary.
- However, the derivative $\frac{d}{dx} \text{Var}_{\tilde{x} \overset{\text{uni}}{\sim} B_\delta(x)} P(Y = +1 \mid \tilde{x})$ is negative when δ is a small positive quantity. This suggests that the local variance in the model's posterior probability, denoted as $\text{Var}_{\tilde{x} \overset{\text{uni}}{\sim} B_\delta(x)} P(Y = +1 \mid \tilde{x})$, decreases as x diverges from the decision boundary.

Drawing from this analysis, it can be inferred that for the Bayes' optimal classifier, *zones exhibiting higher posterior probabilities are associated with lower local variance*. A significant implication of this finding is that for a trained model, if an input x shows a high posterior probability (indicating high confidence), and minor modifications to x lead to substantial changes in the posterior probability, then it is likely that the model's performance at this input x is suboptimal, suggesting that the model's prediction at this point could be erroneous. This crucial insight inspires us to employ both posterior probability and local variance as criteria in the pseudo-label selection process.

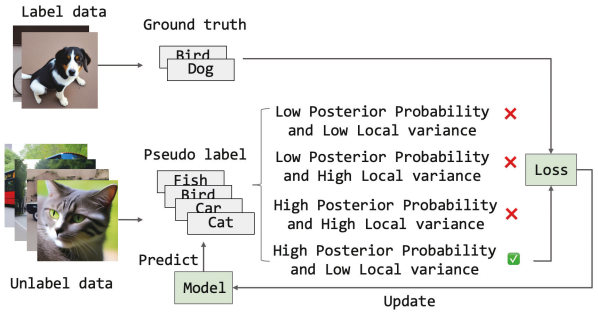

Figure 2: Illustration of our Local Variance Match (LVM) method. For pseudo-labels, we select those with high posterior probability and low local variance to participate in the loss calculation.

Local Variance Match Method

This section introduces the Local Variance Match (LVM) method, a novel strategy for pseudo-label selection in semi-supervised learning (SSL) tasks. The main idea behind LVM is to use local variance as an indicator of pseudo-label accuracy. By adopting this method, we aim to enhance the quality of selected pseudo-labels, which in turn augments the model's performance.

In SSL, the training dataset comprises both labeled and unlabeled samples. Same as the notation in (Wang et al. 2023; Xie et al. 2020; Zhang et al. 2021), we represent the labeled data by $\mathcal{D}_L = \{(x_l, y_l) : l \in [N_L]\}$, where N_L is the number of labeled samples, and y_l stands for the corresponding label of the input x_l. Unlabeled data is symbolized by $\mathcal{D}_U$, and we have $\mathcal{D}_U = \{x_u : u \in [N_U]\}$, with N_U signifying the count of unlabeled samples. The cross-entropy loss is represented as $\mathcal{H}(\cdot, \cdot)$, weak data augmentation (or identity mapping) by $\omega(\cdot)$, and strong data augmentation by $\Omega(\cdot)$. The neural network model is given by $f(\cdot, \theta)$, with θ denoting the trainable parameters.

The supervised loss for labeled data is expressed as:

$$\mathcal{L}_l = \frac{1}{N_L} \sum_{l=1}^{N_L} \mathcal{H}\left(y_l, f\left(\omega\left(x_l\right), \theta\right)\right),$$

Unsupervised training objectives for unlabeled samples, such as those suggested in (Wang et al. 2023; Xie et al. 2020; Zhang et al. 2021), are described as:

$$\begin{aligned} \mathcal{L}_u = \frac{1}{N_U} \sum_{u=1}^{N_U} &\mathbb{1}\left[\max\left(f\left(\omega\left(x_u\right), \theta\right)\right) > \tau\right] \\ &\times \mathcal{H}\left(f\left(\omega\left(x_u\right), \theta\right), f\left(\Omega\left(x_u\right), \theta\right)\right) \end{aligned}$$

Inspired by our theoretical insights, we propose the inclusion of local variance (LV) during the pseudo-label selection phase. The local variance of model f at data point x is defined by:

$$LV_f(x) = Var_{\tilde{x} \overset{uni}{\sim} B_\delta(x)}(f(\tilde{x}))_{i_x}, \tag{9}$$

Here, i_x denotes the i_x-th component of the probability vector, with i_x representing the predicted category of the neural network f for the original input x. Specifically,

$$i_x = \text{argmax}_i f(x)_i. \tag{10}$$

In practice, to compute the local variance, approximations to Equation 9 are necessary. Here, we employ a direct approximation method: we introduce noise perturbations to the input x to produce $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \ldots, \tilde{x}_k$. To reduce computational load, it is possible to set $\tilde{x}_1 = x$. After forwarding through the neural network f, we obtain k probabilities: $f(\tilde{x}_1)_{i_x}, f(\tilde{x}_2)_{i_x}, \ldots, f(\tilde{x}_k)_{i_x}$. We then compute the statistical variance between these k probabilities to approximate $LV_f(x)$. Mathematically, we express this as:

$$LV_f(x) \approx \frac{1}{k-1} \sum_{j=1}^{k} \left(f(\tilde{x_j}))_{i_x} - \frac{1}{k} \sum_{j=1}^{k} f(\tilde{x_j}))_{i_x} \right)^2 \tag{11}$$

In our LVM method, before using unlabeled data, we filter valid pseudo-labels $\mathcal{D}'_U$ based on local variance:

$$\mathcal{D}'_U = \{x \in \mathcal{D}_U | LV(f(\omega(x_u), \theta)) < \tau_1 \wedge \max(f(\omega(x_u), \theta)) > \tau_2\}$$

As such, our unsupervised training objective for unlabeled data, incorporating local variance, is as follows. A visual illustration of our approach is provided in Figure 2.

$$\mathcal{L}_u = \frac{1}{N_U} \sum_{u=1}^{N_U} \mathbb{1}[x_u \in \mathcal{D}'_U] \times \mathcal{H}(f(\omega(x_u), \theta), f(\Omega(x_u), \theta))$$

To summarize, Algorithm 1 incorporates local variance (LV) into the pseudo-label selection criterion. The LVM algorithm enables us to select pseudo-labels that not only possess high posterior probabilities (above τ_2) but also exhibit low local variance (below τ_1). This unique perspective of considering local variance helps address the limitations of existing SSL methods that rely solely on the absolute value of posterior probabilities for pseudo-label selection. The local variance allows us to effectively identify and exclude noisy or mispredicted labels.

Algorithm 1: One iteration of Local Variance Match (LVM) Algorithm

Require: Labeled data $\mathcal{D}_L$, Unlabeled data $\mathcal{D}_U$, Local variance threshold τ_1, Posterior probability threshold τ_2, Unlabeled data weight λ_u

1: Compute labeled data loss:

$$\mathcal{L}_l = \frac{1}{N_L} \sum_{l=1}^{N_L} \mathcal{H}(y_l, f(\omega(x_l), \theta))$$

2: Select valid pseudo label with:

$$\mathcal{D}'_U = \{x \in \mathcal{D}_U | LV(f(\omega(x_u), \theta)) < \tau_1 \wedge \max(f(\omega(x_u), \theta)) > \tau_2\}$$

3: Compute unlabeled data loss:

$$\mathcal{L}_u = \frac{1}{N_U} \sum_{u=1}^{N_U} \mathbb{1}[x_u \in \mathcal{D}'_U] \times \mathcal{H}(f(\omega(x_u), \theta), f(\Omega(x_u), \theta))$$

4: Compute total loss: $\mathcal{L} = \mathcal{L}_l + \lambda_u \cdot \mathcal{L}_u$

5: Update model parameters θ using the computed total loss $\mathcal{L}$

Experiments

Image Classification Experiments

In this subsection, we aim to evaluate the efficacy of the Local Variance Match (LVM) method for image classification tasks.

Experimental Setup We assess the performance of our Local Variance Match (LVM) approach in image classification across several benchmark datasets: CIFAR-10, CIFAR-100, SVHN, and ImageNet (Russakovsky et al. 2015), adjusting the quantity of labeled data for each. Specifically, our experiments involve 40, 250, and 4000 labeled data points for CIFAR-10; 2500 labeled data points for CIFAR-100; 40 and 250 labeled data points for SVHN; and 100,000 labeled data points for ImageNet. We further test our method under the most stringent conditions, using only 10 labeled data points for CIFAR-10 and 400 labeled data points for CIFAR-100. These settings are widely applied in the comparison of SSL algorithms such as FlexMatch (Zhang et al. 2021), MixMatch (Berthelot et al. 2019), FreeMatch (Wang et al. 2023), and UDA (Xie et al. 2020). For a fair comparison, we maintain the same network architectures and similar training iterations (around 1,000,000) as other benchmarks such as FlexMatch (Zhang et al. 2021), SoftMatch (Chen et al. 2023), and USB (Wang et al. 2022). We employ Wide ResNet-28-2 (Zagoruyko and Komodakis 2017) for CIFAR-10 and SVHN, Wide ResNet-28-8 for CIFAR-100, and ResNet50 for ImageNet, consistent with FreeMatch (Wang et al. 2023) for a fair comparison.

Under the most stringent conditions (10 labeled data points for CIFAR-10 and 400 labeled data points for CIFAR-100), we utilize the SGD optimizer with the following parameters: a learning rate of 0.03, a cosine learning rate decay schedule, momentum of 0.9, and a weight decay of 0.0005. We set τ_1 to a relative value of 0.97, meaning we exclude the top 3% of pseudo-labels with high local variance. For experiments on the large-scale ImageNet, we employ the SGD optimizer with a learning rate of 0.1, a cosine learning rate decay schedule, momentum of 0.9, and a weight decay of 0.0003.

For all other configurations, we utilize the SGD optimizer with the following parameters: a learning rate of 0.03, a cosine learning rate decay strategy, momentum of 0.9, and a

	CIFAR-10			
Label data number	10	40	250	4000
Pseudo Label (Lee et al. 2013)	80.21	74.61	46.49	15.08
MeanTeacher (Tarvainen and Valpola 2017)	76.37	70.09	37.46	8.10
MixMatch (Berthelot et al. 2019)	65.76	36.19	13.63	6.66
ReMixMatch (Berthelot et al. 2020)	20.77	9.88	6.30	4.84
UDA (Xie et al. 2020)	34.53	10.62	5.16	4.29
Dash (Xu et al. 2021)	27.28	8.93	5.16	4.36
FlexMatch (Zhang et al. 2021)	13.85	4.97	4.98	4.19
FreeMatch (Wang et al. 2023)	8.07	4.90	4.88	**4.10**
LVM(Ours)	**5.41**	**4.87**	**4.84**	4.19
Fully-Supervised	4.62			

Table 1: Comparison of error rates (expressed in %) on the CIFAR-10 dataset across various unsupervised learning methods, including Pseudo Label, Mean Teacher, MixMatch, ReMixMatch, UDA, Dash, FlexMatch, FreeMatch, and our proposed method, Local Variance Match (LVM), for different quantities of labeled data (10, 40, 250, 4000). The best-performing results are highlighted in bold. For reference, the performance of a fully-supervised model is also provided. Notably, our LVM method achieves an error rate of 5.41% in the most challenging setting of CIFAR-10 with only 10 labels, which is less than 1% different from the fully-supervised setting.

weight decay of 0.0005. The parameter τ_1 is set to 0.98, resulting in the exclusion of the top 2% of pseudo-labels with high local variance. Additionally, we determine a value for τ_2 aligning with (Wang et al. 2023) to ensure fairness. All experiments were carried out on four Tesla V100 32GB GPUs. In all experiments, we compute the approximation of local variance following equation 11. We choose $k = 2$ or $k = 3$ to minimize computational overhead, and we use Gaussian noise to obtain the perturbations.

Experiment Results The results of our image classification experiments on CIFAR-10, as well as CIFAR-100 and SVHN datasets, are presented in Tables 1 and 2 respectively. Our Local Variance Match (LVM) method consistently achieves state-of-the-art performance across various data configurations, underscoring its effectiveness.

Compared to other unsupervised learning techniques, such as ReMixMatch (Berthelot et al. 2020), UDA (Xie et al. 2020), Dash (Xu et al. 2021), FlexMatch (Zhang et al. 2021), and FreeMatch (Wang et al. 2023), our LVM technique demonstrates superior performance. For instance, in the most stringent CIFAR-10 scenario with only 10 labels (one label per class), LVM records an error rate of 5.41. This significantly outperforms other advanced methods such as FreeMatch, which register an error rate of 8.07. Impressively, the error rate of 5.41 attained by our method is remarkably close to that of fully-supervised learning, which is 4.62. In another stringent CIFAR-100 scenario with only 400 labels (four labels per class), LVM records an error rate of 35.87. This represents a nearly 2% improvement over the previous best results. These results emphasize the potential of our LVM method to excel even in situations with extremely limited labeled data, making it a robust and effective solution for unsupervised learning scenarios where labeled data is either scarce or costly to obtain.

Additionally, to validate the effectiveness of the LVM method on large-scale data, we also conducted experiments on ImageNet with 100k labeled data points. Our LVM method significantly outperforms the latest counterpart by 1.26% on top-1 error and 1.65% on top-5 error. The experimental results are presented in Table 3.

Analysis of Experimental Results To check the functioning of LVM and evaluate the efficacy of our local variance strategy in filtering out inaccurate pseudo labels, we conduct a comprehensive analysis. We use the experiment carried out on CIFAR-10 with 40 labeled data points as a representative case.

We start by examining the mask ratio across different methods. This metric signifies the fraction of pseudo-labels that are discarded. As depicted in Figure 3, the overall mask ratio for LVM is slightly higher than for other methods. This is consistent with our hypothesis, as LVM discards pseudo-labels that exhibit high local variance.

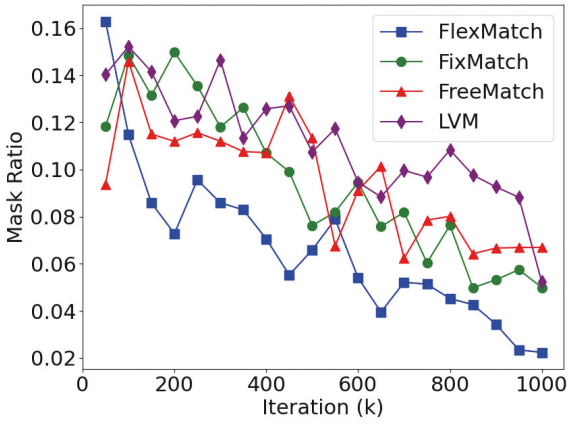

Figure 3: Mask ratio comparison for various methods.

	CIFAR-100		SVHN	
Label data number	400	2500	40	250
Pseudo Label (Lee et al. 2013)	87.45	57.74	64.61	15.59
MeanTeacher (Tarvainen and Valpola 2017)	81.11	45.17	36.09	3.45
MixMatch (Berthelot et al. 2019)	67.59	39.76	30.60	4.56
ReMixMatch (Berthelot et al. 2020)	42.75	**26.03**	24.04	6.36
UDA (Xie et al. 2020)	46.39	27.73	5.12	1.92
Dash (Xu et al. 2021)	44.82	27.15	2.19	2.04
FlexMatch (Zhang et al. 2021)	39.94	26.49	8.19	6.59
FreeMatch (Wang et al. 2023)	37.98	26.47	1.97	1.97
LVM(Ours)	**35.87**	26.52	**1.94**	**1.94**
Fully-Supervised	19.30		2.13	

Table 2: Comparison of error rates (%) on the CIFAR-100 and SVHN datasets for various semi-supervised learning methods with different amounts of labeled data is presented. The number of labeled data used for each experiment is indicated in the first row. The best results are highlighted in bold. Our proposed method, Local Variance Match (LVM), consistently achieves the lowest error rates in most settings, demonstrating its robustness and effectiveness compared to other state-of-the-art methods. For reference, the performance of a fully-supervised model is also provided. Notably, LVM, in the most challenging setting of CIFAR-100 with only 400 labels, achieves an error rate of 35.87%. This marks over a 2% improvement over the previous state-of-the-art results.

Method	top-1	top-5
fixmatch	43.66	21.80
flexmatch	41.85	19.48
freematch	40.57	18.77
LVM(ours)	**39.31**	**17.12**

Table 3: Performance comparison on ImageNet dataset.

To further validate our approach, it's essential to ascertain whether LVM operates as expected:

Does discarding pseudo-labels with high local variance indeed remove more erroneous labels?

This is illustrated in Figure 4. The blue line represents the accuracy of the valid pseudo-labels, while the red line represents the accuracy of the pseudo-labels discarded due to their high local variance. These discarded labels possess high posterior probabilities and, therefore, are not filtered out by the posterior probability threshold. Notably, in order to better reflect the entire training process, we sampled data points at fixed iteration intervals, and a moving average option was utilized in the plotting of Figure 3 and Figure 4.

The results unambiguously indicate that the pseudo-labels discarded based on our local variance criterion possess significantly lower accuracy compared to the overall accuracy of the valid pseudo-labels. This highlights the effectiveness of our local variance approach in filtering pseudo-labels. Our method provides a fresh perspective on pseudo-label selection, differentiating from prior works that predominantly centered around posterior probability, as cited in (Berthelot et al. 2020, 2019; Wang et al. 2023).

Speech Recognition Experiments

In addition to our main experiments, we apply the Local Variance Match (LVM) method to speech recognition tasks to further validate the effectiveness of our proposed approach and to demonstrate the broad applicability of LVM.

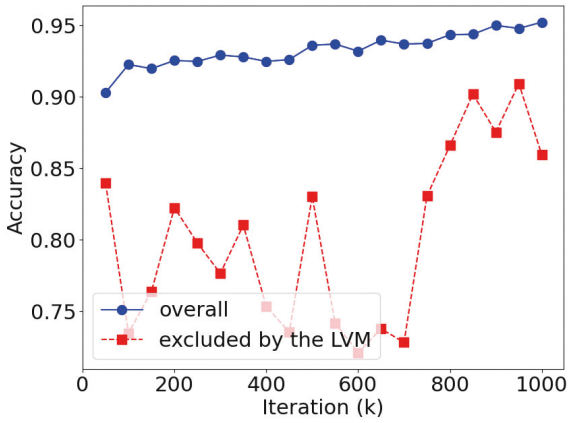

Figure 4: Validation of LVM functionality through pseudo-label accuracy

Experimental Setup For the speech recognition experiment, we use the AISHELL-1 dataset (Bu et al. 2017) (approximately 150 hours) as the labeled data and the AISHELL-2 dataset (Du et al. 2018) (approximately 1000 hours) as the unlabeled data. We evaluate our model based on the Word Error Rate (WER) measured on the AISHELL-1 test set. Given the slower convergence rate of speech models compared to image classification tasks, we adopt an offline pseudo-label usage strategy. Specifically, we use a pre-trained model from WeNet (Yao et al. 2021) on the AISHELL-1 dataset as the pseudo-label generator, generate pseudo-labels on the AISHELL-2 dataset, and concurrently

Supervised	Unsupervised		
Aishell-1	Aishell-2		Result
	Posterior Probability Threshold	Local Variance Threshold	Word Error Rate (%)
All Used	0.03	Off	4.72
All Used	0.03	0.05	**4.72**
All Used	0.03	0.1	4.74
All Used	0.025	Off	4.78
All Used	0.025	0.05	**4.71**
All Used	0.025	0.1	4.77
All Used	0.02	Off	4.77
All Used	0.02	0.05	4.75
All Used	0.02	0.1	**4.74**
All Used	No Unsupervised Data		5.17

Table 4: ASR task results using various thresholds for posterior probability and local variance.

train a new model with the labeled data from AISHELL-1 and selected pseudo-labels from AISHELL-2. For these experiments, we use a state-of-the-art hybrid CTC/attention architecture (Watanabe et al. 2017), incorporating a Conformer (Gulati et al. 2020) encoder, a 2D convolutional input layer, and a Transformer (Vaswani et al. 2017) decoder. We adopt a batch size of 16, an Adam optimizer with a learning rate of 0.002, and a warmup scheduler with 25,000 steps. Gradients are clipped at 5 and accumulated over 4 steps, and we train for a total of 80 epochs for each setting.

Experiment Results Our findings indicate that the incorporation of local variance has allowed us to achieve superior results with less data, further substantiating the efficacy of our Local Variance Match (LVM) approach. As observed in Table 4, under various settings, the implementation of our LVM strategy for refining pseudo-labels consistently enables the model to achieve better (or equivalent) performance with fewer data. For example, the setting with a posterior probability threshold of 0.025 and a local variance threshold of 0.05 used fewer pseudo-labels compared to the setting with only a posterior probability threshold of 0.025, yet it resulted in improved performance.

Related Work

The emergence of pseudo-labeled semi-supervised learning (SSL) techniques has garnered significant attention. The seminal study, Pseudo Label (Lee et al. 2013), treated pseudo-labels as genuine labels during training. Ensuring the quality of pseudo-labels is crucial for the success of SSL. Subsequent advanced SSL methods like UDA (Xie et al. 2020), FixMatch (Sohn et al. 2020), Flexmatch (Zhang et al. 2021), Class-Imbalanced Adaptive Thresholding (Guo and Li 2022), Dash (Xu et al. 2021), Adamatch (Berthelot et al. 2022), and Freematch (Wang et al. 2023) have focused on pseudo-label selection, implementing either fixed or adaptive thresholds based on the model's confidence levels.

UDA (Xie et al. 2020) emphasizes the significance of advanced data augmentation methods in SSL. FixMatch (Sohn et al. 2020) produces pseudo-labels from weakly augmented unlabeled images, retaining only those with high confidence. Flexmatch (Zhang et al. 2021) introduces Curriculum Pseudo Labeling (CPL), adjusting class-specific thresholds at every iteration based on the model's current learning stage. (Rizve et al. 2020) adopts a model-centric stance, drawing insight from calibration. Freematch (Wang et al. 2023) autonomously adjusts the confidence threshold in line with the model's learning progression.

Together, these SSL techniques reveal a range of strategies for pseudo-label selection, leveraging their quality to enhance model performance. However, these methods rely on the absolute magnitude of the model's predicted posterior probabilities for pseudo-label identification. While this strategy is straightforward and intuitive, it poses certain challenges. Specifically, when dealing with sparsely labeled data, the model might mistakenly predict with high confidence, potentially compromising the effectiveness of these methods. Our newly introduced Local Variance Match (LVM) approach addresses these challenges by incorporating local variance as a criterion to assess pseudo-label accuracy. This inclusion establishes a more reliable pseudo-label selection process, leading to improved performance across various tasks.

Conclusion

In this study, we introduced the Local Variance Match (LVM) method, a state-of-the-art technique for pseudo-label selection in semi-supervised learning. By emphasizing local variance, LVM enhances the quality of pseudo-labels, leading to superior model performance on benchmark datasets, such as CIFAR-10, CIFAR-100, and ImageNet. Our method also performs well in speech recognition tasks. In future work, we believe that characterizing the theoretical properties of the LVM algorithm is an important direction.

Acknowledgements

The authors thanks the support of National Key R&D Program of China, under Grant No. 2020AAA0104500.

References

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2020. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. arXiv:1911.09785.

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

Berthelot, D.; Roelofs, R.; Sohn, K.; Carlini, N.; and Kurakin, A. 2022. AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation. arXiv:2106.04732.

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

Bu, H.; Du, J.; Na, X.; Wu, B.; and Zheng, H. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, 1–5. IEEE.

Chan, W.; Jaitly, N.; Le, Q. V.; and Vinyals, O. 2015. Listen, Attend and Spell. arXiv:1508.01211.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. SoftMatch: Addressing the Quantity-Quality Trade-off in Semi-supervised Learning. arXiv:2301.10921.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33: 22243–22255.

Du, J.; Na, X.; Liu, X.; and Bu, H. 2018. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale. arXiv:1808.10583.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; and Pang, R. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. arXiv:2005.08100.

Guo, L.-Z.; and Li, Y.-F. 2022. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International Conference on Machine Learning*, 8082–8094. PMLR.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

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

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2020. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *International Conference on Learning Representations*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. arXiv:1609.03499.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Chen, H.; Fan, Y.; Sun, W.; Tao, R.; Hou, W.; Wang, R.; Yang, L.; Zhou, Z.; Guo, L.-Z.; et al. 2022. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35: 3938–3961.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; Schiele, B.; and Xie, X. 2023. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. arXiv:2205.07246.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; Le, Q.; Agiomyrgiannakis, Y.; Clark, R.; and Saurous, R. A. 2017. Tacotron: Towards End-to-End Speech Synthesis. arXiv:1703.10135.

Watanabe, S.; Hori, T.; Kim, S.; Hershey, J. R.; and Hayashi, T. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8): 1240–1253.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33: 6256–6268.

Xu, Y.; Shang, L.; Ye, J.; Qian, Q.; Li, Y.-F.; Sun, B.; Li, H.; and Jin, R. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, 11525–11536. PMLR.

Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; and Lei, X. 2021. WeNet: Production oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit. arXiv:2102.01547.

Zagoruyko, S.; and Komodakis, N. 2017. Wide Residual Networks. arXiv:1605.07146.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.

Zhu, Z.; Luo, T.; and Liu, Y. 2022. The Rich Get Richer: Disparate Impact of Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*.