# VIBRATION-BASED UNCERTAINTY ESTIMATION FOR LEARNING FROM LIMITED SUPERVISION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We investigate the problem of estimating uncertainty for training data, so that deep neural networks can make use of the results for learning from limited supervision. However, neither the prediction probability nor the entropy can accurately capture the uncertainty of out-of-distribution data. In this paper, we present a novel approach that measures the uncertainty from the vibration of sequential data, *e.g.*, the output probability during the training procedure. The key observation is that, a training sample that suffers heavier vibration often offers richer information when it is manually labeled. We make use of the Fourier Transformation to measure the extent of vibration, deriving a powerful tool that can be used for semi-supervised, active learning, and one-bit supervision. Experiments on the CIFAR10, CIFAR100, and mini-ImageNet datasets validate the effectiveness of our approach.

## 1 INTRODUCTION

Recently deep learning (LeCun et al., 2015) has become the main methodology for the computer vision tasks. However, training deep neural network usually needs tremendous labeled data which costs amounts of labors. Researchers have proposed some approaches (Rasmus et al., 2015; Grandvalet et al., 2005; Han et al., 2021; Luo et al., 2013) for learning from limited supervision, including semi-supervised learning and active learning. All of them aim to utilize the large amounts of unlabeled data to improve the model training. Hence, obtaining an accurate estimation to the predictive uncertainty for unlabeled data is quite important. The existing uncertainty estimated methods, *e.g.*, the predictive probabilities (Lewis & Gale, 1994) and the entropy (Wang et al., 2016), usually fail in estimating the uncertainty for the out-of-distribution data. We argue that they cannot reflect the real model uncertainty. From another aspect, Bayesian methods offer a natural probabilistic representation of uncertainty in deep learning. We develop a Gaussian approximation to connect the model optimization to the Bayesian procedure by sampling from the latter part of the training procedure.

We propose a novel approach that estimates the predictive uncertainty using the sampled sequential information. A series of predictive probabilities for each unlabeled sample can be obtained by a forward pass after each training epoch. To satisfy the requirement of Bayesian approximation, the sequence is formed by using the latter training epochs. We consider to estimate the uncertainty by measuring the vibration of this sequence. The description to vibration contains two keys: (i) where the baseline it fluctuates around, and (ii) how large is its fluctuations. This inspires us to utilize the Fourier Transformation (FT) to measure it, in which the direct component represents the fluctuation baseline, and the high frequency parts reflect the fluctuation degree. Combining these two parts we will obtain the accurate estimation to the uncertainty. Notably here the usage to FT is independent of the order of the sequence, which is consistent with the sampling theory. To further improve this approach, we consider to combine it with the label flipping information. By fusing them with a weight parameter, a more accurate uncertainty estimation will be obtained. As show in Figure 1, the instantaneous probabilities often provide inaccurate estimation to uncertainty, *e.g.*, the image with high probability and high vibration owns a wrong prediction. Our approach that takes an overall consideration to the fluctuation baseline and intensity will avoid this bother.

We apply this uncertainty estimation approach to improve the model that learning from limited supervision, *e.g.*, semi-supervised learning, active learning and the recently proposed one-bit supervision (Hu et al., 2020). Semi-supervised approaches roughly can be categorised into two types,

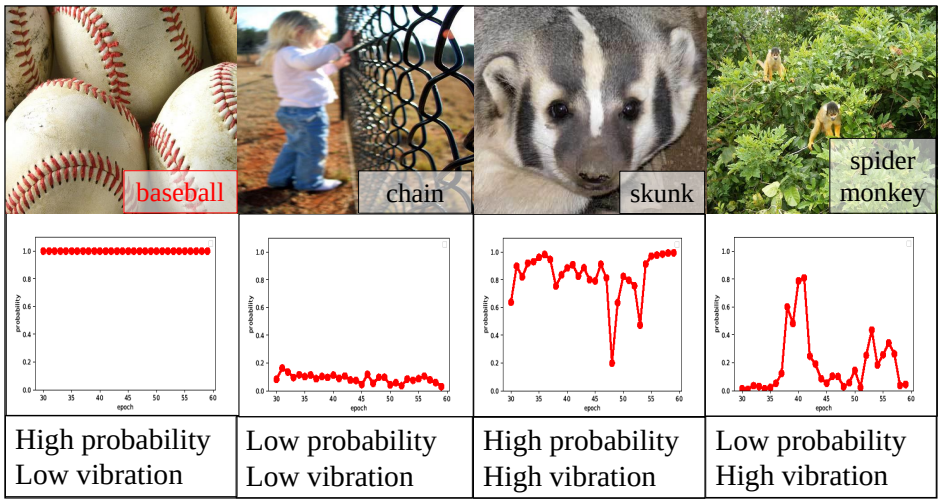



Figure 1: The four types of selected samples, including the images and their corresponding scatter diagram of probabilities sequence. The textboxes on the images represent their predictive labels, where the red text denotes a correct prediction while the black texts denote incorrect ones. The experiments are conducted on ImageNet trained using $\sim 3\%$ labeled samples.

*i.e.*, the consistency-based approaches and the pseudo-labeling based approaches. Here we show that these two kinds of approaches can be combined by using the latter to improve the former. We use the proposed approach to select the reliable pseudo labels and generate accurate weights for unlabeled samples, to improve the consistency-based baseline. For active learning, we utilize our approach to select and annotate the most uncertain samples and re-train the model, to show its effectiveness in selecting the informative samples. Finally, we apply the proposed approach to one-bit supervision, a recently proposed weakly supervised algorithm which annotates an image by answering whether it belongs to a specific class. To improve this method, we propose a mix annotation approach which combining the advantage of full-bit and one-bit annotation. In particular, we select the appropriate samples based on the estimated uncertainty for the two kinds of annotation method to maximize the gains.

We evaluate our approach on CIFAR10, CIFAR100 and Mini-ImageNet for the three tasks. The extensive experiments demonstrate that, the proposed approach enjoys superiority in selecting no matter reliable pseudo labels and the informative samples, and most of all, making an accurate estimation to the uncertainty for each unlabeled data.

## 2 Related Work

**Semi-Supervised Learning** (Rasmus et al., 2015; Laine & Aila, 2016; Miyato et al., 2018) often can be categorised into two types according to their usages of unlabeled data. The first type assigns pseudo labels (Cascante-Bonilla et al., 2020) to the unlabeled data and optimizes them with the labeled data together. (Lee et al., 2013) proposed to use the class with the maximum probability as the true labels for the unlabeled data. (Iscen et al., 2019) improved it by using a transductive label propagation method to obtain more accurate pseudo labels. The second type utilizes the consistency regularization (Kuo et al., 2020; Han et al., 2021) to facilitate the model training. The methods to form the consistency are various, *e.g.*, Mean Teacher (Tarvainen & Valpola, 2017) inputted the sample with different perturbations into the teacher and student model respectively to make their outputs similar. WCP (Zhang & Qi, 2020) imposed additive noise on network weights and making structural changes. In addition, some methods aim to combine these two types of approaches, *e.g.*, MixMatch (Berthelot et al., 2019b) introduced a single loss to seamlessly reduce the entropy while maintaining consistency. ReMixMatch (Berthelot et al., 2019a) improved it by introducing two new techniques of distribution alignment and augmentation anchoring.

**Active Learning** aims to select informative samples to annotate to reduce the labeling cost. According to the selection criterion it can be classified into two groups. Firstly, the diversity-based methods select samples that can represent the whole distribution of the unlabeled pool, *e.g.*, (Shi & Yu, 2019) proposed to identify a small number of samples that best represent the overall data space. (Sener & Savarese, 2017) proposed to choose a subset that minimizes a bound between an average loss over it and the remaining data. (Sinha et al., 2019) utilized the variational autoencoder and adversarial network to choose samples that are not well represented in the labeled set. The second type utilizes the uncertainty (Ash et al., 2019) to select samples that can decrease the model uncertainty, *e.g.*, using the probability of a predicted class (Lewis & Gale, 1994), the entropy of the class posterior probabilities (Wang et al., 2016), and the target losses (Yoo & Kweon, 2019). BALD (Houlsby et al., 2011) chose data points that are expected to maximise the mutual information between predictions and model posterior. BatchBALD (Kirsch et al., 2019) selected informative samples utilizing a tractable approximation to the mutual information between a batch of samples and model parameters.

**Uncertainty Estimated Approaches** usually are used to select the informative data for active learning, *e.g.*, (Gao et al., 2020) used the consistency-based metric for selecting uncertain samples. (Huang et al., 2021) did this by evaluating the discrepancy of outputs of different optimization steps which can be used to estimate the sample loss. They can also be used for other tasks, *e.g.*, UPS (Rizve et al., 2020) used the MC Dropout (Gal & Ghahramani, 2016) method to measure the predictive uncertainty to selected reliable pseudo labels. AUM (Pleiss et al., 2020) utilized the average difference between the logit values for a sample's assigned class and its highest non-assigned class to identify the mislabeled data.

# 3 APPROACH

In this section, we introduce the proposed approach that estimates the predictive uncertainty from the view of sequential data. Here we firstly introduce the setting of learning from limited supervision, and show the significance of acquiring uncertainty for this task. Then we elaborate the proposed approach to provide a solution for this. Finally, to verify its effectiveness we apply our approach to three tasks of learning from limited supervision, namely, semi-supervised learning, active learning and one-bit supervision.

## 3.1 LEARNING FROM LIMITED SUPERVISION

For the setting of learning from limited supervision, we often have a dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}$, where $\mathbf{x}_n$ is the $n$-th sample of image data and $N$ is the total number of training samples. Let $y_n^\star$ denote the ground-truth class label of $\mathbf{x}_n$ with $C$ kinds of values, and they are mostly unseen in our setting. An initial set of samples $S^0$ is chose randomly to partition the dataset into two subsets $\mathcal{D}^{\mathrm{S}}$ and $\mathcal{D}^{\mathrm{U}}$, where the superscripts respectively represent 'supervised' and 'unsupervised'. Learning from limited supervision aims to utilize the unlabeled data to reduce the model uncertainty. Hence, we write the objective as,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}^{\mathrm{S}}}\ell(\mathbf{y}_n^\star,\ \mathbf{f}(\mathbf{x};\boldsymbol{\theta})) + \lambda\cdot\mathbb{E}_{\mathbf{x}\in\mathcal{D}^{\mathrm{U}}}\mathbf{h}\left(q,\ \mathbf{f}(\mathbf{x};\boldsymbol{\theta})\right), \tag{1}$$

where $\mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ represents the model function and $\boldsymbol{\theta}$ is the learnable parameters. The $\ell(\cdot,\cdot)$ is the cross-entropy loss for the labeled samples. The $\mathbf{h}\left(\cdot,\cdot\right)$ denotes the corresponding loss function and $q$ is the obtained information from the unlabeled data using semi-supervised methods or active learning methods. Since the main idea for learning from limited supervision is the utilization of unlabeled data, measuring the uncertainty to distinguish each of them is very significant for this task. Hence, it is necessary to develop an approach to accurately estimate the uncertainty.

## 3.2 UNCERTAINTY ESTIMATION

The conventional measures, *e.g.*, the maximum predictive probabilities, the entropy and the gradients, are often estimated using the instantaneous information. We argue that they cannot measure the real uncertainty for their failure in dealing with the out-of-distribution data, *i.e.*, they might predict a high probability with the wrong label for an outlier. However, Bayesian probability theory provides us a mathematical tool to analysis model uncertainty. The predictive distribution for a Bayesian

procedure is defined as:

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int p(y_*|\theta, \mathbf{x}_*)p(\theta|\mathcal{D})d\theta, \tag{2}$$

where $x_*$ and $y_*$ are test inputs and outputs. The posterior distribution $p(\theta|\mathcal{D})$ in equation 2 is intractable. Next we will show how to develop a Gaussian approximation to the posterior from SGD iterations. According to the deduction in (Mandt et al., 2017), when the gradients or the learning rates are small enough and the optimization is confined to a small enough region, the iterations of stochastic gradient descent equal to a stochastic process, namely Ornstein-Uhlenbeck process (Uhlenbeck & Ornstein, 1930). The OU process has an analytic stationary distribution $q(\theta)$ which is Gaussian:

$$q(\theta) \propto \exp\left\{-\frac{1}{2}\theta^\top \Sigma^{-1}\theta\right\}, \tag{3}$$

where $\Sigma$ is the corresponding covariance matrix. We can approximate $q(\theta)$ by the Monte Carlo sampling procedure, *e.g.*, drawing $\theta$ from the latter part of the training procedure. Then we use this Gaussian distribution to approximate the posterior $p(\theta|\mathcal{D})$, to make it possible to describe the predictive distribution.

In the following, we introduce how to estimate the uncertainty based on the above theory. Here we give a general description to this question. Supposing we have the sequence $\mathbf{e} = \{e_i\}_{i=0}^{R}$ where $R$ is the length, and it is the results sampled from the satisfactory training epochs. We aim to utilize this sequential data to estimate the predictive uncertainty for each unlabeled data. To achieve this goal, we consider to calculate the vibration of this sequence. In general, if the sequence has higher vibration intensity around a lower baseline, it represents the prediction is more uncertain. This inspires us to utilize the Fourier Transformation to capture the vibration. In the following, we will introduce the technical details for our approach.

### 3.3 Vibration-based Approach

Now we we get back to the tasks of learning from limited supervision. Firstly we train an initial model for $T$ epochs in a semi-supervised type, *e.g.*, the Mean Teacher algorithm. We select the latter part of the training procedure, *e.g.*, starting from the $M$-th epoch and ending at the $L$-th epoch (where the model converges). Then by conducting forward pass at the corresponding training epoch, we can obtain the sequence of outputs $\left\{\mathbf{y}_n^M, \mathbf{y}_n^{M+1}, \ldots, \mathbf{y}_n^L\right\}$, where $\mathbf{y}_n^i$ is the $C$-dimension vector for $n$-th sample of $i$-th epoch. To better describe the vibration, we re-form the sequence by using the probabilities of a specified class, *i.e.*, defining the sequence as $\mathbf{s}_n = \left\{s_n^i\right\}_{i=M}^{i=L}$ where $s_n^i$ is the $c$-th element of $\mathbf{y}_n^i$ and $c$ is the class with maximum probability predicted in $L$-th epoch. We utilize the Fourier Transformation to estimate the uncertainty from the frequency domain, which is denoted as

$$\mathbf{S}_k = \mathscr{L}\{\mathbf{s}^n\} = \sum_{i=M}^{L} s_n^i \cdot e^{-j\frac{2\pi}{L-M}ki}. \tag{4}$$

By calculating the real part of $\mathbf{S}_k$, we obtain the amplitude values $\{A_0, A_1, \ldots, A_{L-M+1}\}$ for the corresponding frequency components. Because of the conjugate symmetry of Discrete Fourier Transformation, we just use the half part of the obtained amplitudes $\left\{A_0, \cdots, A_{(L-M+1)/2}\right\}$ (assuming it is even). The $A_0$ represents the direct component of the frequency, which reveals the baseline where the sequence fluctuates. And the $\left\{A_1, \ldots, A_{(L-M+1)/2}\right\}$ represents the high frequency part which tells the vibration intensity. To avoid the negative effect of the noise, we drop a few components with the highest frequency after $P$-th amplitude. Therefore, we define the predictive uncertainty using the sequence information by

$$v_c = \sum_{i=1}^{P} A_i - \mu \cdot A_0, \tag{5}$$

where $\mu$ is the weight coefficient for balancing the high frequency parts and the direct component. Here the summation to the high frequency parts lets the sequence lose its order, which makes it no conflict with the sampling theory. We believe that utilizing equation 5 to estimate the uncertainty will be more accurate. In addition, we consider that the other kind of information within the outputs,

namely the label flipping, is also useful for estimating the uncertainty. Generally, a prediction is more uncertain when the predicted label flips more frequent in the training process. Hence, we define a sequence with binary values $\{b_M, b_{M+1}, \ldots, b_L\}$ for each unlabeled sample, where $b_i = 1$ denotes $\arg\max \mathbf{y}_n^i$ is equal to $\arg\max \mathbf{y}_n^L$ and $b_i = 0$ denotes they are different.

For the label flipping sequence, we also conduct Discrete Fourier Transformation on it and calculate its vibration according to equation 5, which denoted as $v_l$. To conveniently combine these two measures, we conduct min-max normalization for them to obtain the results $\hat{v}_c$ and $\hat{v}_l$ respectively. Finally, we define the fused predictive uncertainty by a weight parameter $\alpha$ as:

$$v_f = (1 - \alpha) \cdot \hat{v}_c + \alpha \cdot \hat{v}_l \tag{6}$$

Our approach can make more accurate estimation to the uncertainty than the methods using instantaneous probabilities or the entropy, which can be verified by the experiments in section 4.2. We attribute this to the approach of calculating vibration for the sequential data. Several methods also attempted to utilize the dynamic information to improve the model training. (Zhou et al., 2020b) introduced the approach of dynamic instance hardness to guide curriculum learning procedure. (Zhou et al., 2020a) extended this approach to train with noisy labels. Our approach is different from them in three respects, firstly, they conducted on the labeled data while our approach is applied to the unlabeled data. Secondly, they calculated the hardness by exponential moving average of an instantaneous measure (*e.g.*, the losses) over time, while we use the Fourier Transformation to capture the vibration of the sequential data.

### 3.4 Applying to Learning from Limited Supervision

In this section, we introduce to apply the proposed approach to the tasks of learning from limited supervision, *e.g.*, semi-supervised learning and active learning, including the recently proposed one-bit supervision. We argue that the difference among these tasks is the usage to the unlabeled data. Semi-supervised learning utilizes the initial labeled samples and the amounts of unlabeled samples to train the model. By the usage of unlabeled data we can classify it into two groups, namely, the consistency-based and the pseudo-labeling based methods. The former often utilizes a consistency loss to optimize the unlabeled samples, which equals to define the $h(\cdot, \cdot)$ in equation 1 as a mean square error loss and set $q$ as the outputs of samples with perturbation. The latter often utilizes the model trained on the labeled set to generate pseudo labels for unlabeled samples, which equals to set $q$ to the pseudo labels and uses a cross-entropy loss to optimize them.

Comparing to semi-supervised learning, except for the initial labeled set, active learning also select informative samples to annotate to improve the performance of learning from limited supervision. The objective of active learning is equal to set $h(\cdot, \cdot)$ in equation 1 as the cross entropy and let $q$ be the ground-truth labels of the selected samples. Different from active learning which annotates the true label for each image, one-bit supervision uses a weakly annotation method which annotates an image by asking the labeler if it belongs to a guessing class. It improves the tasks of learning from limited supervision by proposing a new weakly annotation method. All of the three tasks start by training the initial model using $\mathcal{D}^{\mathrm{S}}$ and $\mathcal{D}^{\mathrm{U}}$. In the following, we will introduce these applications in detail.

**Semi-Supervised Learning**. For semi-supervised learning, we consider that the consistency-based methods still can benefit from the accurate pseudo labels. In particular, after obtaining the initial model $\mathbb{M}_0$, we calculate the uncertainty for each unlabeled sample according to equation 6, by using the two kinds of sequential data, namely, the series of probabilities and label flipping outputted by $\mathbb{M}_0$. Then, we select $K$ samples with the smallest vibration values from $\mathcal{D}^{\mathrm{U}}$ and use $\mathbb{M}_0$ to generate pseudo labels for them. Then adding the selected samples to $\mathcal{D}^{\mathrm{S}}$ and fine-tuning the model to $\mathbb{M}_1$ by equation 1 using all the information. In the next stage, we moderately increase the number of selected samples to obtain more accurate pseudo labels. The cycle of selecting certain samples and fine-tuning the model will continue until the training converges. Because the selected samples are with the least uncertainty, we believe that these generated reliable pseudo labels will be accurate enough to improve its consistency-based baseline.

Besides selecting certain samples to obtain their pseudo labels, we also assign weights for the unlabeled data in each stage. The consistency-based methods often treat all the unlabeled data equally, *i.e.*, each of the sample share the same weight in the unlabeled loss. In general, the samples with different uncertainty will have different impacts to the training. And the more emphasis on the highly

uncertain samples will hurt the training. Hence, we consider to use our uncertainty estimation approach to generate weights for each unlabeled sample. The weights is defined as:

$$w = 1 - \frac{v_f - \min v_f}{\max v_f - \min v_f}. \tag{7}$$

**Active Learning**. We apply the proposed uncertainty estimation approach to active learning by using it to select the most uncertain samples. The training process of active learning often consists of several iterations and each of them annotates a batch of selected samples. In the cycle $t$, we utilize $\mathbb{M}_{t-1}$ to estimate the uncertainty for the unlabeled data according to equation 5 and equation 6. Here we use both the single measure and the fused measure for active learning to verify the superiority of the latter one. Then selecting $J$ samples with the largest values of $v_f$ to check their ground-truth, and adding them to $\mathcal{D}_{t-1}^{\mathrm{S}}$ by removing them from $\mathcal{D}_{t-1}^{\mathrm{U}}$. Then we update the model to $\mathbb{M}_t$ using both $\mathcal{D}_t^{\mathrm{S}}$ and $\mathcal{D}_t^{\mathrm{U}}$.

**One-bit Supervision**. Though better performance has been achieved, one-bit supervision still has its limitations, *i.e.*, it fails in obtaining the correct labels of the most uncertain samples. To address this issue, we propose a mix annotation approach which combines the full-bit and one-bit annotation approach. Notably one-bit supervision maintains a budget of supervision and partitions it into several stages to use. The training process is similar to that of active learning. To avoid repetition, we only introduce the difference between them. For the $t$-th stage, after calculating the uncertainty estimation for the unlabeled samples according to equation 6, we select $I$ samples with the largest vibration values to conduct full-bit annotation, then adding them to $\mathcal{D}^{\mathrm{F}}$, the subset of full-bit annotated samples. Next, we select a subset $\mathcal{D}_t^{\mathrm{O}}$ from $\mathcal{D}_t^{\mathrm{U}}$ and use $\mathbb{M}_{t-1}$ to make predictions for them to conduct one-bit annotation. We know that conducting one-bit annotation for the samples with the predictive probabilities around $0.5$ will achieve the highest gains. Hence, we select the middle-uncertain samples, *i.e.*, the samples around the boundary of the accuracy sorted by the obtained uncertainty. By checking the ground-truth, we add the correctly predicted samples to the set of right guesses $\mathcal{D}^{\mathrm{O}+}$, and the incorrectly predicted ones are added to the wrong guessing set $\mathcal{D}^{\mathrm{O}-}$. Then combining the positively labeled set $\mathcal{D}^{\mathrm{S}} \cup \mathcal{D}^{\mathrm{F}} \cup \mathcal{D}^{\mathrm{O}+}$, the negatively labeled set $\mathcal{D}^{\mathrm{O}-}$ and the unlabeled set $\mathcal{D}_t^{\mathrm{U}}$ to update the model.

# 4 Experiments

## 4.1 Datasets and Implementation Details

**Dataset**. For semi-supervised learning and active learning, we do experiments on three classification benchmarks CIFAR10, CIFAR100 (Krizhevsky et al., 2009) and Mini-ImageNet. CIFAR10 and CIFAR100 are standard datasets with 10 and 100 classes respectively. They both contains 60K images in which 50K for training and 10K for testing. All of them are $32 \times 32$ RGB images and uniformly distributed over all classes. For Mini-ImageNet, we use the training/testing split created in (Ravi & Larochelle, 2016), which contains 100 classes, 50K training images and 10K testing images. For one-bit supervision, the experiments are conducted on CIFAR100 and Mini-ImageNet.

**Implementation details**. In order to better compared with other methods, we use the Wide ResNet-28-2 (Zagoruyko & Komodakis, 2016) as the backbone for CIFAR10 and CIFAR100, and use the ResNet-18 (He et al., 2016) for Mini-ImageNet, for the experiments of *semi-supervised learning*. For *active learning*, the ResNet-18 is used as the backbone for the three datasets. For *one-bit supervision*, we follow the experimental setting in (Hu et al., 2020) by using ResNet-50 for Mini-ImageNet, and a 26-layer deep residual network (He et al., 2016) with Shake-Shake regularization (Gastaldi (2017)) for CIFAR100. All the experiments are conducted based on the Mean Teacher algorithm (Tarvainen & Valpola, 2017). The experiments are all trained for 180 epochs in each stage, and the consistency parameter is set to 10, 1,000 and 100 respectively. The balance coefficient $\mu$ is set to $0.1$ for all experiments. The fused weight $\alpha$ is set to $0.2$ for CIFAR10 ($\alpha = 0.8$ for the first two active learning cycles) and CIFAR100, and $0.5$ for Mini-ImageNet.

For *semi-supervised learning*, we set the number of selected pseudo labels $K$ to $20\%$, $40\%$ and $60\%$ of the training set for the experiments on CIFAR10 using 1000, 2000 and 4000 labels respectively, then increasing it by $5\%$ in the next cycle. For CIFAR100, the $K$ is set to $20\%$ and $40\%$ respectively for experiments using 4000 and 10000 labels, and we increase it by $2\%$ each cycle. The setting of

Table 1: Test error (%) of semi-supervised methods on CIFAR10, CIFAR100 and Mini-ImageNet. The methods with $*$ represent that using the CNN-13 architecture. For our method and the baseline Mean Teacher, we report the mean and standard deviation over 3 runs.

| | CIFAR10 | | | CIFAR100 | | Mini-ImageNet |
|---|---|---|---|---|---|---|
| Total Labels | 1000 | 2000 | 4000 | 4000 | 10000 | 10000 |
| PL (Lee et al., 2013) | 30.91±1.73 | 21.96±0.42 | 16.21±0.11 | - | 36.21±0.19 | - |
| TSSDL* (Shi et al., 2018) | 21.13±1.17 | - | 10.90±0.23 | - | - | - |
| DeepLP* (Iscen et al., 2019) | 22.02±0.88 | - | 12.69±0.29 | 46.20±0.76 | 38.43±1.88 | 57.58±1.47 |
| Π model (Laine & Aila, 2016) | - | - | 14.01±0.38 | - | 37.88±0.11 | - |
| VAT (Miyato et al., 2018) | 18.64±0.40 | 14.40±0.15 | 11.05±0.31 | - | - | - |
| MT (Tarvainen & Valpola, 2017) | 21.54±0.12 | 15.59±0.95 | 11.48±0.21 | 52.36±0.39 | 38.00±0.17 | 56.91±0.16 |
| Ours | 16.94±0.18 | 12.09±0.60 | 9.33±0.08 | 46.56±0.43 | 34.55±0.21 | 54.91±0.08 |

Mini-ImageNet is similar to that for CIFAR100. We only run 4 cycles for each experiment which is far smaller than (Rizve et al., 2020). For *active learning* on CIFAR10, we first randomly select 100 samples as the initial labeled set, and add 500 samples in each of the following stage, except for the last one which we add 1000 samples. For CIFAR100, we initially select 5000 samples randomly as the labeled set and add 1000 samples in the next stage. For Mini-ImageNet, we randomly select 20% samples as the initial set and add 5% in the following stage. We compare three commonly used samples selecting approaches here. "Random" represents random selection. "Confidence" indicates utilizing the maximum predictive probabilities to select. "K-center" (Sener & Savarese, 2017) is a diversity-based method which selects samples by maximizing the distance between the candidates and its nearest neighbours. For *one-bit supervision*, we split the quota of supervision used in each stage into two parts, 1000 full-bit annotations (about 6644 bits of supervision) and the remaining one-bit annotations.

### 4.2 Main Results

**Semi-Supervised Learning**. The experiments are conducted on CIFAR10, CIFAR100 and Mini-ImageNet. Expect for the main baseline Mean Teacher, we also list several semi-supervised methods including the pseudo-labeling based methods PL (Lee et al., 2013), TSSDL (Shi et al., 2018), and DeepLP (Iscen et al., 2019), and the consistency-based methods Π model (Laine & Aila, 2016) and VAT (Miyato et al., 2018). As shown in Table 1, our approach achieves higher performance than most of the listed consistency-based and pseudo-labeling based methods. In particular, comparing to the baseline Mean Teacher, it brings 4.63%, 3.50% and 2.15% accuracy gains on CIFAR10 respectively for the experiments using 1000, 2000 and 4000 labels. For CIFAR100, it achieves 5.80%, 3.45% gains for the experiments using 4000, 10000 labels respectively. The accuracy gain is 2.00% for that using 10000 labels on Mini-IMageNet. This results show that our approach brings obvious improvements for semi-supervised learning by selecting reliable pseudo labels and weighting unlabeled samples. We attribute this to the accurate estimation of the predictive uncertainty for all the unlabeled samples. As Figure 2 shows, the proposed approach achieves higher accuracy of the pseudo labels for selecting the same percentage of samples, than the predictive probabilities, the entropy and the consistency methods. It verifies that our approach selects more accurate pseudo labels to improve the consistency-based

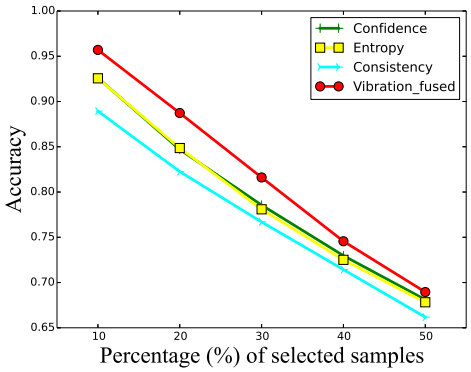
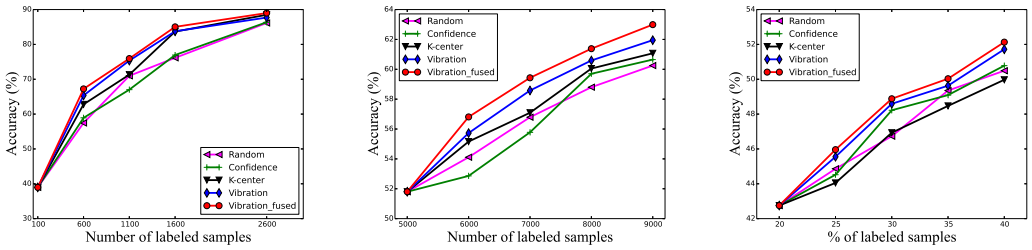



Figure 2: The accuracy of the selected pseudo labels by four kinds of methods, namely confidence, entropy, consistency and the fused vibration measure. The "Consistency" is calculated by the L2 distance between the two kinds of augmentations on one image.

semi-supervised baseline. Notably, the proposed uncertainty estimation approach can be combined with more powerful semi-supervised model and achieve better performance.




(a) CIFAR10 (b) CIFAR100 (c) Mini-ImageNet

Figure 3: Active learning test accuracy on CIFAR10, CIFAR100 and Mini-ImageNet. The comparison methods are random selection, the maximal predictive paobabilities, and the k-center approach, as well as the vibration measure using probability sequence.

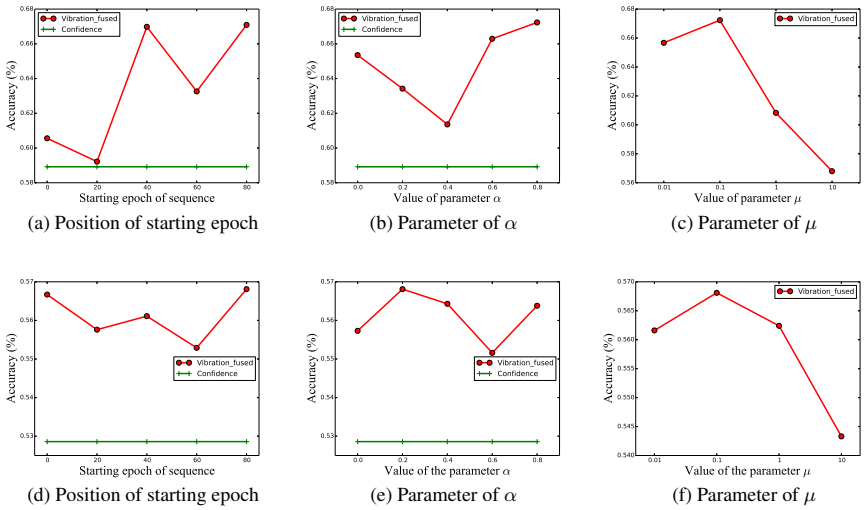
**Active Learning**. To verify the effectiveness of our approach on active learning, we do experiments on CIFAR10, CIFAR100 and Mini-ImageNet. From the results in Figure 3 we can obtain some observations. Firstly, the approach using the probabilities sequence only (denoted by "Vibration") achieves higher performance than all the baselines in most cases on three datasets. For example, with 600 labels on CIFAR10, it outperforms "Confidence" and "K-center" respectively by $\sim 6.4\%$ and $\sim 2.6\%$. With 6000 labels on CIFAR100, the accuracy gains become $\sim 2.9\%$ and $\sim 0.6\%$ when compared to the "Confidence" and "K-center" method. This shows that our approach selects more informative samples than the two baselines, especially the former using instantaneous probabilities. Secondly, the fused approach further improves the performance on three datasets. In particular, compared to the "Vibration" approach, it brings $\sim 1.9\%$ and $\sim 1.1\%$ accuracy gains respectively on CIFAR10 with 600 labels and CIFAR100 with 6000 labels. These results demonstrate the effectiveness of our approach that incorporating the label flipping information to estimate the more accurate uncertainty. In addition, our approach can be integrated easily with other structure of networks and more powerful semi-supervised algorithms, which reveals its potential for the tasks of learning from limited supervision.

**One-bit Supervision**. We also use the proposed approach to improve one-bit supervision, a weakly annotation version of active learning. The experiments are conducted on CIFAR100 and Mini-ImageNet. The proposed mix annotation approach achieves $\mathbf{76.07}\%$ and $\mathbf{49.71}\%$ accuracy respectively on the two datasets. When compared to the semi-supervised baseline Mean Teacher, it brings $6.31\%$ and $8.65\%$ accuracy gains respectively on CIFAR100 and Mini-ImageNet. Mix annotation still achieves $2.31\%$ and $4.17\%$ gains respectively compared to one-bit supervision. In addition, the MixMatch (Berthelot et al., 2019b) and UDA (Xie et al., 2019) achieves $74.12$ and $75.50$ accuracy on CIFAR100 using Wide ResNet-28-8 backbone, which is inferior to our approach. Such significant improvements reveal that the proposed approach accurately captures the uncertainty for the unlabeled data. And, the mix annotation approach which combines the full-bit and one-bit annotation maximizes the labeling gains.

### 4.3 Analysis

**Transfer to large scale dataset**. To verify the effectiveness of our approach on the large scale dataset, we do experiments on ImageNet (Russakovsky et al., 2015) for semi-supervised learning. The experiments are based on the ResNet-50 backbone. The accuracy for the experiments using $5\%$ and $10\%$ labels respectively is $52.70\%$ and $59.88\%$. Comparing to our baseline Mean Teacher, the proposed approach brings $4.89\%$ and $2.87\%$ accuracy gains respectively for the experiments trained with $5\%$ and $10\%$ labels. This obvious improvements reveal its potential of applying to the large scale dataset. We also argue that the uncertainty estimation approaches enjoy the superiority of generalizing to large scale datasets for learning from limited supervision tasks.

**Position of starting epoch**. We do experiments on CIFAR10 and CIFAR100 for the first active learning stage to investigate the effect of starting epoch. The results are shown in the subfigure (a)




(a) Position of starting epoch (b) Parameter of $\alpha$ (c) Parameter of $\mu$

(d) Position of starting epoch (e) Parameter of $\alpha$ (f) Parameter of $\mu$

Figure 4: Analysis to the sequence position, balance coefficient $\mu$ and fused weight $\alpha$. The first row lists the results of CIFAR10, and the second row lists the results of CIFAR100. Experiments are conducted on the first iteration of active learning on the two datasets.

and (d) in Figure 4. We can observe that the overall performance of our approach is superior to the "Confidence" method on both two datasets, and the best results are achieved when setting the starting point to the latter training epoch. This is in accord with our conduct that sampling from the latter part of training procedure to approximate the posterior.

**Robustness to Hyperparameters**. Our approach introduces two hyperparameters, namely the balance coefficient $\mu$ and the fused coefficient $\alpha$. Here we investigate the effect of these parameters bring to our approach. The experiments are conducted on the first active learning stage on CIFAR10 and CIFAR100. The parameter $\mu$ plays the role of balancing the high frequency part and the direct component. As shown in the subfigure (c) and (f) of Figure 4, we can observe that $\mu = 0.1$ achieves the best performance for both two datasets. It shows the importance to make a balance between the vibration baseline and its intensity. We set $\mu = 0.1$ for all the experiments, which shows the robustness of our approach. The parameter $\alpha$ is used to balance the two components in the fused measure. As shown in the subfigure (b) and (e) of Figure 4, we tune the $\alpha$ from $0.0$ to $0.8$. One can observe that all the parameter settings achieve higher performance than the "Confidence" method. This indicates that our approach is robust to this parameter. Setting $\alpha$ to $0.8$ and $0.2$ achieves the best accuracy respectively on on CIFAR10 and CIFAR100. By these analysis, we can conclude that the proposed approach enjoys flexibility in hyperparameter adjustment, for the limited hyperparameter numbers and the robustness to them.

## 5 Conclusions

In this paper we propose a novel approach that estimates the uncertainty by calculating the vibration of the sequential data, and use it to improve learning from limited supervision. We argue that neither the predictive probabilities nor the entropy can capture the accurate uncertainty because they usually fail in dealing with the out-of-distribution data. Inspired by the Bayesian theory which provides a natural probabilistic representation of uncertainty, we consider to approximate the predictive distribution by sampling from the optimization iterations. After obtaining the sampled sequence data, *e.g.*, the probabilities, we measure its vibration using Fourier Transformation. By being equipped with label flipping, we obtain a more accurate estimation which can be applied to semi-supervised learning and active learning. The effectiveness of the proposed approach is verified by the extensive experiments on CIFAR10, CIFAR100 and Mini-ImageNet. It is believed that the proposed approach can be be integrated with stronger semi-supervised algorithms to achieve better performance, and in the future we will investigate to extend it to other vision tasks, *e.g.*, semantic segmentation.

## REPRODUCIBILITY STATEMENT

Here we show the efforts we make to strengthen the reproducibility of this work. Firstly, we give the clear reference for the theory used in our work, *e.g.*, in Section 3.2. Secondly, the technique detail of the proposed approach is introduced in Section 3.3 and Section 3.4, to make sure the clear understanding to others. Thirdly, we introduce the used datasets and the network structure, as well as the specific experimental settings in Section 4.1 in considerable detail. This gives guarantee to reproduce our experiments for the related researchers. Finally, we also promise to release the corresponding code after we clear up it in a few weeks. This let the researchers who are interested in our work develop their own works based on our approach.

## ETHICS STATEMENT

This paper presents a new uncertainty estimation approach for learning from limited supervision. We summarize the potential impact of our work in the following aspects.

- **To training with limited supervision.** It is an urgent requirement to extract knowledge from unlabeled data. Our work offers a new method to estimate the uncertainty for unlabeled data, which can obviously improve the tasks of learning from limited supervision. We believe that it can be generalized to other vision tasks after follow-up efforts.
- **To the community.** The proposed approach estimates the predictive uncertainty from the sequential view. This gives the community a new inspiration that defining the measures using the sequence data can achieve better performance. We believe the study on these problems can advance the research community.
- **To the society.** The debate on the impact that AI can bring to the human society is long-lasting. Our method has the potential to generalize the existing AI algorithms to more applications, while it also raises the concern of privacy. Therefore, our work can bring both beneficial and harmful impacts and it depends on the motivation of the users.

## REFERENCES

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019b.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pp. 510–526. Springer, 2020.

Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005.

Tao Han, Wei-Wei Tu, and Yu-Feng Li. Explanation consistency training: Facilitating consistency-based semi-supervised learning with interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7639–7646, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Hengtong Hu, Lingxi Xie, Zewei Du, Richang Hong, and Qi Tian. One-bit supervision for image classification. *Advances in Neural Information Processing Systems*, 33, 2020.

Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. *arXiv preprint arXiv:2107.14153*, 2021.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5070–5079, 2019.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32:7026–7037, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, pp. 479–495. Springer, 2020.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

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

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

Wenjie Luo, Alex Schwing, and Raquel Urtasun. Latent structured active learning. *Advances in Neural Information Processing Systems*, 26:728–736, 2013.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Geoff Pleiss, Tianyi Zhang, Ethan R Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*, 2020.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*, 2015.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

Weishi Shi and Qi Yu. Integrating bayesian and discriminative sparse kernel machines for multi-class active learning. *Advances in neural information processing systems*, 2019.

Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 299–315, 2018.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Liheng Zhang and Guo-Jun Qi. Wcp: Worst-case perturbations for semi-supervised deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3912–3921, 2020.

Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: From clean label detection to noisy label self-correction. In *International Conference on Learning Representations*, 2020a.

Tianyi Zhou, Shengjie Wang, and Jeff A Bilmes. Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, 33, 2020b.