Vibration-based Uncertainty Estimation for Learning from Limited Supervision

Hengtong Hu[1], Lingxi Xie[3], Xinyue Huo[3], Richang Hong[1,2⋆], and Qi Tian[3⋆]

[1] School of Computer Science and Information Engineering,
Hefei University of Technology,
[2] Institute of Data Space, Hefei Comprehensive National Science Center,
[3] Huawei Inc

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, both prediction probability and entropy estimate uncertainty from the instantaneous information. In this paper, we present a novel approach that measures 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. Motivated by Bayesian theory, we sample the sequences from the latter part of training. 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, mini-ImageNet and ImageNet datasets validate the effectiveness of our approach.

Keywords: Uncertainty estimation, semi-supervised learning, active learning, one-bit supervision

1 Introduction

Recently deep learning [28] 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 for learning from limited supervision, including semi-supervised learning [40, 15, 27] and active learning [32, 17, 10, 38]. 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 [30] and the entropy [52], usually estimate uncertainty using the instantaneous information, and achieves unsatisfied performance. We aim to utilize the sequential information from training procedure to obtain a more accurate estimation.

⋆ Corresponding authors. Email:hongrc@hfut.edu.cn, tian.qi1@huawei.com.

In general, a series of predictive probabilities for unlabeled samples can be obtained by a forward pass after each training epoch, and we use the probabilities of the class predicted by the last epoch model. We consider to estimate uncertainty by measuring the vibration of this sequence. The description to vibration consists of two keys: (i) where the baseline it fluctuates around, and (ii) how large are its fluctuations. This inspires us to utilize the Fourier Transformation (FT) to measure it. The direct component of its results represents the fluctuation baseline, while the high frequency parts reflect the fluctuation degree. By combining the two parts, an accurate estimation of the uncertainty will be obtained. To further improve this measure, we equip it with the label flipping information, in which each element indicates whether the label changed.

To obtain the appropriate sequence from training process, we utilize Bayesian methods [36, 37] which offer a natural probabilistic representation of uncertainty in deep learning. By sampling from the latter training epochs we connect the model optimization with the Bayesian procedure, which offers theory foundation for our approach. As shown in Figure 1, the instantaneous probabilities might provide inaccurate estimation to uncertainty, *e.g.*, the images with high probability and high vibration own wrong predictions. Our approach that considers both the fluctuation baseline and intensity will alleviate this issue.

We develop methods to apply this uncertainty measure to the tasks of learning from limited supervision, *e.g.*, semi-supervised learning, active learning and the recently proposed one-bit supervision [19]. To improve SSL, we use the proposed measure to select reliable pseudo labels, to further improve the semi-supervised baselines. Also a strategy of class weights is utilized to alleviate the class imbalance issue. Different from SSL that directly utilizes unlabeled data to enhance generalization ability, active learning aims to select informative samples from the unlabeled set to annotate. This is also an appropriate application scenario for our approach. We select and annotate the highly uncertain samples to conduct active learning to verify its effectiveness in uncertainty estimation. Finally, we propose a mix annotation approach to improve one-bit supervision. It utilized a weakly annotation method to efficiently utilize the supervision information, while only negative labels can be obtained for the most uncertain samples. Hence, we propose to incorporate full-bit annotation with one-bit annotation, *i.e.*, using the proposed approach to select appropriate samples to conduct this two kinds of annotation respectively.

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

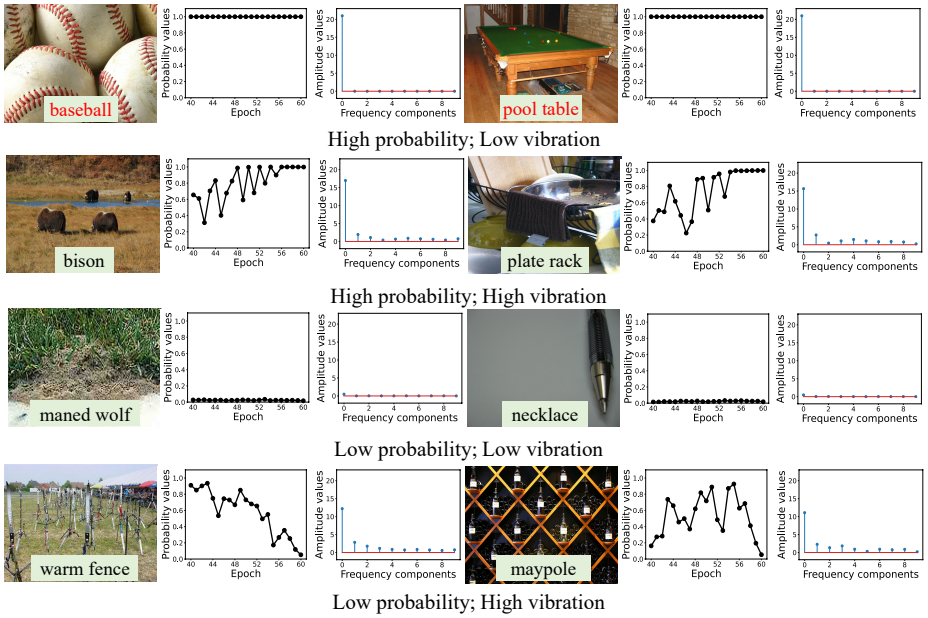


Fig. 1: The four types of selected samples. The first and fourth columns are the images, the second and fifth columns are their corresponding scatter diagram of probabilities sequence, and the third and sixth columns are their magnitude spectra. The textboxes on the images represent their predictive labels, where the red text denotes a correct prediction while the black texts denote incorrect ones. The experiments are conducted on ImageNet trained using $\sim 3\%$ labels

2 Related Work

2.1 Semi-Supervised Learning

Semi-supervised learning [40, 27, 34] often can be categorised into two types according to their usages of unlabeled data. The first type assigns pseudo labels [29, 6] to unlabeled data and optimizes them with labeled data together. Iscen *et al.* [22] used the transductive label propagation method to obtain more accurate pseudo labels. Hu *et al.* [20] proposed a pair loss to minimize the distance between high confidence pseudo labels. The second type utilizes the consistency regularization [26, 17] to facilitate model training. The methods of encouraging the consistency are various, *e.g.*, Mean Teacher [48] inputted a sample with different perturbations into two models to make their outputs be similar. WCP [56] imposed additive noise on network weights and making structural changes. In addition, some methods aim to combine two types of approaches, *e.g.*, MixMatch [4] introduced a single loss to seamlessly reduce the entropy while maintaining consistency. ReMixMatch [3] improved it by extra introducing distribution alignment and augmentation anchoring.

2.2 Active Learning

Active learning aims to reduce labeling cost by selecting informative samples to annotate. According to the selection criterion it can be classified into two groups. Firstly, the diversity-based methods [44] select samples that can represent the whole distribution of the unlabeled pool, *e.g.*, Shi *et al.* [45] proposed to identify a small number of samples that best represent the overall data space. Sinha *et al.* [46] utilized the variational autoencoder and adversarial network to choose samples that are not well represented in the labeled set. The second type utilizes uncertainty [2] to select samples that can decrease the model uncertainty, *e.g.*, using the prediction probability [30], the entropy [52], and the target losses [54]. Gao *et al.* [12] used the consistency-based metric for selecting uncertain samples. Huang *et al.* [21] did this by evaluating the discrepancy of outputs of different optimization steps.

2.3 Uncertainty Estimation Approaches

Bayesian neural networks usually are used to estimate uncertainty, while they are inefficient and computationally intractable. Then some approximated Bayesian inference methods [5, 31] were proposed to alleviate this. Gal *et al.* [11] proposed to estimate uncertainty by interpreting dropout neural networks as variational Bayes. The similar approaches include SpatialDropout [49] and DropBlock [14]. SDE-Net [24] proposed to quantify uncertainty from a dynamical system perspective. AUM [39] 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

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 $\mathbf{x}_n$ and C is the number of classes, and they are mostly unseen in the 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 unlabeled data to reduce model uncertainty. Therefore, we write the objective as,

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

where $\mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ represents the model function and $\boldsymbol{\theta}$ is the learnable parameters. The $\ell(\cdot,\cdot)$ is cross-entropy loss for labeled samples. The $\mathbf{h}\left(\cdot,\cdot\right)$ denotes the loss that utilizes unlabeled data by q, which is obtained via the semi-supervised or active learning methods. Since the main idea for this task is the use strategy

of unlabeled data, measuring uncertainty to distinguish each of them is very significant. Hence, it is necessary to develop an accurate uncertainty estimation approach for learning from limited supervision.

3.2 Vibration-based Approach

The conventional measures, *e.g.*, the maximum predictive probabilities, the entropy and the gradients, often used instantaneous information to estimate uncertainty. We do this from another view, *i.e.*, evaluating it using information from training procedure. Supposing an initial model is trained for T epochs in a semi-supervised type, *e.g.*, the Mean Teacher algorithm. If we sample the model weights from the training process, *e.g.*, from M-th epoch to L-th epoch, and conducting forward pass at each epoch, a sequence of outputs $\{\mathbf{y}_n^M, \mathbf{y}_n^{M+1}, \ldots, \mathbf{y}_n^L\}$ can be obtained, where $\mathbf{y}_n^i$ is the C-dimension vector for n-th sample of i-th epoch. To better describe vibration, we form the sequence as $\mathbf{s}_n = \{s_n^i\}_{i=M}^{i=L}$ 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 aim to utilize this sequential information to estimate uncertainty for unlabeled data. To achieve this, we consider to calculate vibration for this sequence. In general, if the sequence has higher vibration intensity around a lower baseline, the prediction will be more uncertain. This inspires us to utilize Fourier Transformation to capture its vibration. It is denoted as

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

By calculating the real part of $\mathbf{S}_k$, the amplitude sequence $\{A_0, A_1, \ldots, A_{L-M+1}\}$ can be obtained for the corresponding frequency components. Because of the conjugate symmetry of Discrete Fourier Transformation, we use half part of the obtained amplitudes $\{A_0, \cdots, A_{(L-M+1)/2}\}$. A_0 represents the direct component of the frequency, which reveals the baseline where the sequence fluctuates, and $\{A_1, \ldots, A_{(L-M+1)/2}\}$ represents the high frequency part which tells the vibration intensity. Therefore, we define the predictive uncertainty by

$$v_c = \sum_{i=1}^{(L-M+1)/2} A_i - \mu \cdot A_0, \tag{3}$$

where μ is the weight coefficient for balancing high frequency parts and direct component. The summation to the high frequency parts lets the sequence lose its order and makes it have no conflict with the sampling theory. While there may other methods to extract uncertainty, we believe that Eq. (3) is a straightforward and effective method that combines the direct and high frequency parts of the amplitudes. In addition, another kind of information within the outputs, namely label flipping, is also useful for estimating 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 $\arg\max \mathbf{y}_n^i$ equals to $\arg\max \mathbf{y}_n^L$ and $b_i = 0$ represents they are different.

For the label flipping sequence, we also conduct Discrete Fourier Transformation to it and calculate its vibration according to Eq. (3), which denoted as v_l. To conveniently combine the two measures, we conduct min-max normalization for them and obtain the results $\hat{v}_c$ and $\hat{v}_l$ respectively. We verify the effectiveness of this fused measure by the experiments on active learning in Section 4.2. Finally, the predictive uncertainty is defined by a weight α as:

$$v_f = (1 - \alpha) \cdot \hat{v}_c + \alpha \cdot \hat{v}_l \tag{4}$$

3.3 Theoretical Foundation

Since we estimate uncertainty by using sequential information from training process, the important question to be solved is how to sample the sequence. One can use the whole sequence (from 0-th epoch to the last) or part of it. To solve this, we make use of Bayesian probability theory which provides a mathematical tool to analysis model uncertainty. The predictive distribution for a Bayesian procedure is defined as:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\boldsymbol{\theta}, \mathbf{x}_*) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}, \tag{5}$$

where $\mathbf{x}_*$ and $\mathbf{y}_*$ are test inputs and outputs. The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ in Eq. (5) is intractable. Next we will show how to develop a Gaussian approximation to the posterior by stochastic gradient descent (SGD) iterations. According to the deduction in [33], when the gradients or the learning rates are small enough and the optimization is confined to a sufficiently small region, the SGD iterations is equivalent to a stochastic process known as the Ornstein-Uhlenbeck (OU) process [51]. The OU process has an analytic stationary distribution $q(\boldsymbol{\theta})$ which follows a Gaussian distribution of:

$$q(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right\}, \tag{6}$$

where $\boldsymbol{\Sigma}$ is the corresponding covariance matrix. We can approximate $q(\boldsymbol{\theta})$ by Monte Carlo sampling procedure, *e.g.*, drawing $\boldsymbol{\theta}$ from the latter part of the training procedure. To approximate the $p(\boldsymbol{\theta}|\mathcal{D})$, the variational inference [23] is utilized by minimizing the KL divergence between it and the stationary distribution $q(\boldsymbol{\theta})$, which is written as $\arg\min_{\epsilon,S} \mathrm{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})\right)$. It involves with the learning rate ϵ and mini-batch size S. The learning rate to satisfy this is $\epsilon = 2\frac{S}{N}\frac{D}{\mathrm{Tr}(\mathbf{B}\mathbf{B}^T)}$, where D is the dimension of $\boldsymbol{\theta}$ and $\mathbf{B}\mathbf{B}^T = \mathbf{C}$ is the gradient noise covariance. For an explicit deduction, please kindly refer to [33]. To achieve the requirement of minimization, the learning rate needs to be a small value, and the norm of gradients needs to be small but larger than zero. These conditions are satisfied when we sample $\boldsymbol{\theta}$ from the late training epochs (close

to converging). Hence, the approximated predictive distribution is given by:

$$q(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\boldsymbol{\theta}, \mathbf{x}_*) q(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{7}$$

Since the optimization in the late training epochs can agree with Bayesian procedure, we estimate uncertainty by using the sampled sequence information. This can be viewed as sampling from the distribution $q(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D})$.

Relationship to previous works. Similarly, Temporal Ensembling [27] utilized sequential outputs to obtain weights or predictions. AUM [39] proposed to exploits differences in the training dynamics of clean and mislabeled samples. Though they both aim to make use of the outputs in training epochs, our approach is different from them in many aspects. TE only conducted self-ensembling to obtain more accurate predictions, while our approach estimates uncertainty for unlabeled samples. AUM used the front part outputs in training to calculate area under the margin values, which is inapplicable to unlabeled samples for the margins are very close at those epochs. Also, AUM neglects the semantic changes in training epochs while our approach considers this by calculating v_l. In addition, the idea of forgetting events [50] is similar to the used label flipping sequence. Lastly, SG-MCMC [7, 9] is also a Bayesian method that used for uncertainty estimation. However, Our approach is different from them in theory, *i.e.*, our approach utilizes a stationary distribution to approximate the posterior, while SG-MCMC samples from an asymptotically exact posterior.

3.4 Application to Different Scenarios of Learning from Limited Supervision

In this section, we apply the proposed approach to the tasks of learning from limited supervision, including semi-supervised learning, active learning, and one-bit supervision. The difference among them is their usage to unlabeled data. SSL utilizes unlabeled data to enhance generalization ability. In particular, the consistency-based type defines the $h(\cdot, \cdot)$ in Eq. (1) as a mean square error loss and sets q as the outputs for the perturbative samples. The pseudo-labeling based type sets q as the pseudo labels and uses a cross-entropy loss to optimize them. Compared to SSL, AL selects informative samples from unlabeled data to annotate. Its objective equals to set $h(\cdot, \cdot)$ as the cross entropy and let q be the ground-truth of the selected samples. Different from AL that annotates true labels for samples, one-bit supervision annotates by asking the labeler if it belongs to a guessed class. All of the three tasks start by training the initial model using $\mathcal{D}^{\mathrm{S}}$ and $\mathcal{D}^{\mathrm{U}}$.

Semi-Supervised Learning. For SSL, we utilize the proposed uncertainty measurement to mine reliable pseudo labels to improve semi-supervised baselines. In particular, after training the initial model $\mathbb{M}_0$, we calculate uncertainty for unlabeled samples by Eq. (4) via the outputted probabilities and label flipping information. 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 them

to $\mathcal{D}^{\mathrm{S}}$ and fine-tuning the model to obtain the first stage model $\mathbb{M}_1$. We adopt appropriate strategies to utilize the mined pseudo labels for the used baseline. For Mean Teacher [48], we inject the pseudo labels according to each unlabeled batch. For FixMatch [47], we replace the pseudo labels generated by the weak-augmented images with ours.

In the next iteration, we moderately increase the number of selected samples to obtain more reliable pseudo labels. The cycle of selecting pseudo labels and fine-tuning continues until the training converges. One issue for pseudo-label selection is about class imbalance, *i.e.*, the correct predictions may focus on partial classes. Especially when the model is not strong enough, this issue will be more obvious. To alleviate this, we assign weight for each class by $w_i = \frac{\mathrm{A}}{\mathbf{n}_i}$, where A is the average number of samples in each class in labeled set (including pseudo labels), and $\mathbf{n}_i$ denotes the actual number of samples in i-th class. Here we normalize the weights by $\hat{w} = w / \max(w)$.

Active Learning. We apply the proposed approach to AL by selecting the informative samples. The training process of AL often consists of several iterations, and in each of them a batch of samples is selected for annotating. To obtain more accurate estimation, we train the model in each stage in semi-supervised type. In our algorithm, for the t-th cycle, we estimate uncertainty for unlabeled data by Eq. (3) and Eq. (4) according to the model $\mathbb{M}_{t-1}$. Then selecting J samples with the largest uncertainty and checking their ground-truth to imitate the process of annotating. Then adding them to $\mathcal{D}^{\mathrm{S}}_{t-1}$, and removing from $\mathcal{D}^{\mathrm{U}}_{t-1}$. Finally we update to obtain new model $\mathbb{M}_t$ using both $\mathcal{D}^{\mathrm{S}}_t$ and $\mathcal{D}^{\mathrm{U}}_t$. This iteration continues until the satisfied performance is achieved.

One-bit Supervision. We apply our approach to one-bit supervision via conducting mix annotation, which efficiently acquires supervision by combining full-bit and one-bit annotation. Since the multi-stage training framework in one-bit supervision is similar to the process of AL, we omit the introduction to the iterations. For t-th stage, after calculating uncertainty for unlabeled samples by Eq. (4), we select I samples with the largest vibrations 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 $\mathcal{D}^{\mathrm{O}}_t$ from $\mathcal{D}^{\mathrm{U}}_t$ and use the model $\mathbb{M}_{t-1}$ to make predictions for them to conduct one-bit annotation. Generally, selecting samples with predictive probabilities around 0.5 will obtain the highest gains for one-bit annotation. Hence, the middle-uncertain samples are selected according to the model precision. By checking their ground-truth, we add correctly guesses to the positively labeled set $\mathcal{D}^{\mathrm{O}+}$, and add incorrectly guesses to the negatively labeled set $\mathcal{D}^{\mathrm{O}-}$. Finally, we retrain the model by combining the 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 $\mathcal{D}^{\mathrm{U}}_t$.

4 Experiments

4.1 Datasets and Implementation Details

Dataset. For both SSL and AL, we do experiments on three classification benchmarks CIFAR10, CIFAR100 [25] and Mini-ImageNet. CIFAR10 and CIFAR100

are standard datasets with 10 and 100 classes respectively. They contain 60K images in which 50K for training and 10K for testing. All of them are 32×32 RGB images and uniformly distributed over all classes. For Mini-ImageNet, we use the training/testing split created in [41], which contains 100 classes, 50K training images and 10K testing images. For one-bit supervision, the experiments are conducted on CIFAR100 and Mini-ImageNet. In addition, the experiments on ImageNet [43] are also conducted for SSL. This dataset contains 1.2M images from 1000 classes.

Implementation Details. For SSL, we use Wide ResNet-28-2 [55], a commonly used backbone for CIFAR10 and CIFAR100, and ResNet-18 [18] for Mini-ImageNet. For AL, WRN-28-2 is used for all three datasets. For one-bit supervision, we follow the experimental setting in [19] to use ResNet-50 for Mini-ImageNet, and ResNet-26 [18] with Shake-Shake regularization [13] for CIFAR100. The SSL experiments are based on two famous baselines, Mean Teacher [48] and FixMatch [47]. The experiments for AL are conducted using Mean Teacher. We refer to their original paper to set our parameters. We use the SGD optimizer with momentum. For the hyper-parameters in our approach, we set the balance coefficient μ to 0.1 for all experiments. The fused weight α is set to 0.6 for CIFAR10 and Mini-ImageNet, and 0.2 for CIFAR100. For the value of K in SSL, we choose it according the model precision. We follow the general rules to set the parameter I in AL. Specifically, on CIFAR10, we randomly select 100 samples as the initial labeled set, and add 500 samples in each of the following stage, except for the last two which 1000 samples are added; on CIFAR100, we randomly select 5000 samples as the initial labeled set and add 1000 samples in the next stage; on Mini-ImageNet, we randomly select 20% samples as the initial set and add 5% in the following stage. 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 results for combining with two semi-supervised baselines are shown in Table 1. We run 5 iterations for three datasets for their performance approaches to converge. We can observe that our approach both achieves higher performance when applied to Mean Teacher [48] and FixMatch [47]. Also, more accuracy gains are obtained when combined with the MT algorithm, *e.g.*, it achieves 7.00% gains for CIFAR100 with 4000 labels. Except for the used weak baseline, we also own this to that our approach is more applicable to consistency-based approaches. Meanwhile, our approach still achieves 0.82% gains when combined with FixMatch. Specifically, to our knowledge, we report the best results on Mini-ImageNet both with 4000 and 10000 labels when using ResNet-18. We also list some popular semi-supervised methods in Table 1, and among them UPS [42] is most similar with our approach. It also used an uncertainty estimation method (MC Dropout [11]) to mine pseudo labels. The results show that our approach outperforms UPS both with the two baselines, which verifies the superiority of the proposed uncertainty measure.

Table 1: Test error (%) of semi-supervised methods on CIFAR10, CIFAR100 and Mini-ImageNet. The methods with $*$ represent that using the CNN-13 architecture. "RA" represents the Randaugment [8] approach. For our method and two baselines Mean Teacher and FixMatch, we report the mean and standard deviation over 3 runs

	CIFAR10			CIFAR100			Mini-ImageNet	
Total Labels	250	1000	4000	2500	4000	10000	4000	10000
PL [29]	49.78±0.43	30.91±1.73	16.09±0.28	-	-	36.21±0.19	-	-
DeepLP [22]	-	22.02±0.88*	12.69±0.29*	-	46.20±0.76*	38.43±1.88*	70.29±0.81	57.58±1.47
Π model [27]	-	-	14.01±0.38	-	-	37.88±0.11	-	-
VAT [34]	-	18.64±0.40	11.05±0.31	-	-	-	-	-
PLCB [1]	24.81±5.35	-	6.28±0.30	-	37.55±1.09*	32.15±0.50*	56.49±0.51	46.08±0.11
MixMatch [4]	11.29±0.75	-	6.24±0.07	39.70±0.27	-	28.59±0.31	49.79±0.11	44.27±0.23
UPS [42] (RA)	-	8.18±0.15*	6.39±0.02*	-	40.77±0.10*	32.00±0.49*	-	-
SemCo [35]	5.87±0.31	-	4.43±0.01	33.80±0.57	29.40±0.18	25.07±0.04	46.01±0.93	41.25±0.76
MT [48]	52.30±0.95	21.54±0.12	11.48±0.21	-	52.36±0.39	38.00±0.17	70.58±0.37	56.91±0.16
Ours+MT	48.05±1.17	16.94±0.18	9.33±0.08	-	46.56±0.43	34.55±0.21	69.46±0.13	54.91±0.08
MT (RA)	16.50±0.18	11.72±0.10	9.48±0.29	49.83±0.10	43.86±0.56	35.60±0.36	61.97±0.32	52.98±0.27
Ours+MT (RA)	10.37±0.53	7.63±0.64	5.87±0.05	42.12±0.22	36.86±0.46	29.84±0.23	57.02±0.26	49.73±0.29
FixMatch [47] (RA)	6.16±0.79	5.21±0.08	4.73±0.03	34.28±0.23	31.22±0.16	26.87±0.05	40.02±0.35	38.47±0.39
Ours+FM (RA)	**6.06±0.76**	**4.84±0.04**	**4.63±0.12**	**33.53±0.21**	**30.40±0.20**	**26.18±0.12**	**39.21±0.58**	**37.72±0.20**

To reveal the quality of uncertainty estimated by our approach, we analyze the relationship between it and the Expected Calibration Error (ECE) score[16]. Experiments are conducted on CIFAR10 with 1000 and 4000 labels and CIFAR100 with 4000 and 10000 labels, and trained using MT. The results are presented in Figure 2, which show that the ECE scores are positively associated with the uncertainty values. It means reliable pseudo labels can be obtained by selecting samples with low uncertainty defined in our approach. Experiments on CIFAR100 with 5000 labels are also conducted to verify this. Our approach achieves (95.7%/88.73%) accuracy with top-5,000/10,000 selected samples, while the numbers for Consistency, Confidence, and Entropy are (88.94%/82.26%), (92.6%/84.7%), and (92.56%/84.86%), respectively.

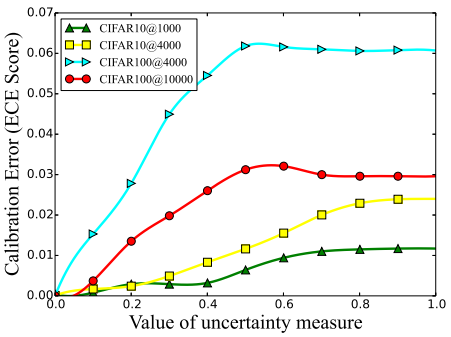


Fig. 2: The relationship between the expected calibration error (ECE) and the value of our uncertainty measure (v_f) on unlabeled samples.

Table 2: Test error (%) of our approach on CIFAR10, CIFAR100, and Mini-ImageNet for active learning. They are all based on the Mean Teacher algorithm. The comparison methods include K-center [44], MC Dropout [11], AUM [39] and Consistency [12]. The Initial labels for these three datasets respectively are 100, 5000 and 10000

Methods		Random	Confidence	K-center	MC Dropout	AUM	Consistency	Vibration	Vibration_fused
CIFAR10	iter 1	37.98	42.39	35.53	33.27	40.75	40.95	35.37	27.33
	iter 2	24.26	23.24	23.20	25.74	25.00	28.38	26.75	20.43
	iter 3	20.85	18.25	16.87	21.46	18.93	18.58	18.79	15.14
	iter 4	15.69	12.80	12.40	13.42	12.28	12.49	11.58	10.53
	iter 5	12.39	10.55	10.70	11.46	10.06	10.05	10.04	9.47
CIFAR100	iter 1	45.49	45.37	45.35	45.31	45.77	46.18	45.64	43.99
	iter 2	43.40	43.15	42.53	42.10	42.37	43.24	41.97	41.22
	iter 3	42.25	41.08	40.03	39.46	39.22	40.30	38.78	38.42
	iter 4	39.67	38.72	38.67	38.01	38.56	38.36	37.61	37.11
	iter 5	38.19	37.37	37.20	37.06	37.33	37.9	37.01	36.51
Mini-ImageNet	iter 1	45.23	46.07	45.13	44.33	45.58	45.00	44.68	43.14
	iter 2	42.24	42.89	42.29	42.43	41.47	42.22	42.41	40.30
	iter 3	40.34	40.95	40.77	40.30	40.07	40.45	38.98	38.62
	iter 4	38.90	39.50	38.82	37.57	38.63	37.43	37.87	36.77
	iter 5	37.51	37.11	37.98	36.55	36.85	36.63	36.96	35.95

Also, with the training iteration increases, uncertainty for unlabeled samples becomes lower. For example, in SSL on CIFAR100 with 10000 labels, the average uncertainty decreases throughout training: the values after 1st, 5th, 10th iterations are 0.3516, 0.2570, and 0.2434, respectively. In addition, we also do ablations for pseudo labels selection and class weights. The experiments are conducted on Mini-ImageNet with 4000 and 10000 labels by using MT. Only mining pseudo labels achieves 42.12% and 49.32% accuracy respectively, and further using the generated class weights brings 0.86% and 0.85% gains. Finally, we test the computation cost on a RTX 2080Ti GPU, in SSL in CIFAR10/100, each training epoch takes 76.72s, in which, after forward/backward prop, using the newest model to update sequential information takes 10.73s, then uncertainty estimation plus pseudo label selection takes 5.44s.

Active Learning. The experiments for AL are conducted on CIFAR10, CIFAR100 and Mini-ImageNet. From the results on Table 2 we can obtain some observations. Firstly, the fused vibration measure achieves higher performance than the single measure on all four iterations for three datasets. This verifies the effectiveness of the approach which utilizes both the outputted probabilities and label flipping sequences. Secondly, compared to the basic AL methods, *e.g.*, Random, Confidence and K-center [44], our approach obviously outperform them in all iterations. For example, on CIFAR10, it achieves 8.20% accuracy

gains compared to K-center in the first iteration. Though these three methods are simple, we argue that they still provide strong baselines when combined with semi-supervised algorithms.

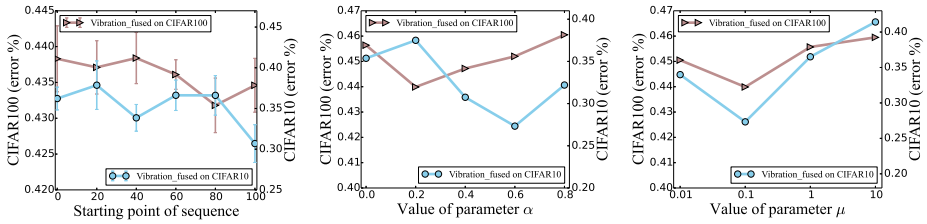
Thirdly, compared to other uncertainty estimation methods, such as MC Dropout [11] and AUM [39], our approach still achieves higher performance in four iterations on three datasets. In particular, the gains are 5.94% and 13.42% respectively in the first iteration on CIFAR10. This demonstrates the superiority of our approach for uncertainty estimation. Here AUM was originally designed to identify the mislabeled data, we adapt it to active learning by calculating the average difference between the prediction probabilities for a sample's pseudo-labeled class and its highest non-pseudo-labeled class. Lastly, on all datasets, the proposed approach outperforms Consistency [12] for all iterations. Notably, it is also designed for SSL algorithm, which is appropriate to verify our approach as a comparison. In addition, another observation on these three datasets is, our approach obtains the highest gains on the first stage, which shows that it enjoys more advantages when the supervision is scarce.

One-bit Supervision. The experiments for one-bit supervision are conducted on CIFAR100 and Mini-ImageNet. The proposed mix annotation approach achieves **23.93**% and **50.29**% test error respectively on these two datasets. Compared to the baseline Mean Teacher, our approach brings 6.31% and 8.65% accuracy gains respectively. The gains still have 2.31% and 4.17% when compared to the original one-bit supervision. In addition, MixMatch [4] and UDA [53] achieves 25.88% and 24.50% test error on CIFAR100 using Wide ResNet-28-8 backbone, which is inferior to our approach. These results demonstrate the effectiveness of the proposed mix annotation approach, for which efficiently utilize the annotation information to maximize the labeling gains. And we own this to that our approach accurately estimates uncertainty for all unlabeled samples.

We also apply the proposed approach to noise learning and conduct experiments on CIFAR10 by assigning random labels to a random subset of training data. Our approach is based on MT and trains using predicted noisy data as unlabeled data. It achieves 10.06%, 12.77% test error for 0.2, 0.4 noise level respectively, which are obviously better than the standard training (using all data) achieves, *i.e.*, 13.82%, 18.19% test error.

4.3 Diagnostic Experiments

Transferring to Large-Scale Dataset. To verify the effectiveness of our approach on large scale datasets, we do experiments on ImageNet [43] for semi-supervised learning. The results are still based on the two baselines, Mean Teacher [48] and FixMatch [47]. For MT, we use ResNet-50 as the backbone. Our approach achieves 47.30% and 59.88% test error respectively for using 5% and 10% labels, and brings 4.89% and 2.87% gains when compared to the baseline. For FixMatch, we use VGG-16 as the backbone and train using 10% labeled samples. The error rate of our approach is 30.84%, which outperforms the baseline by 3.38% gains. These results reveal the potential of the proposed approach in applying to large scale datasets.



(a) (b) (c)

Fig. 3: Analysis to the starting epoch of the sampled sequence, balance coefficient μ and fused weight α. Experiments are conducted on on CIFAR10 and CIFAR100 for the first iteration of active learning, using WRN-28-2 backbone. Results about starting epoch are ran for 10 times.

Position of Starting Epoch. We conduct experiments for the first active learning iteration on CIFAR10 and CIFAR100, to investigate the effect of different starting epoch of the sequence. The results are shown on the subfigure (a) of Figure 3. We can obtain some observations from them. Firstly, most of the results are higher than the "Random" approach on two datasets, which shows the superiority of estimating uncertainty from sequential data. Secondly, the best results are obtained when setting the starting point to the latter epoch, *e.g.*, 80 and 100 for CIFAR10 and CIFAR100 respectively. This is in accord with the sampling theory for approximating the posterior, which is introduced in Section 3.3.

Lastly, we observe that sampling from early epochs does not show dramatic accuracy drop, *i.e.*, violating the constraints necessary for the theoretical guarantee still yield good results. We argue that this is quite often in practice, *e.g.*, for SVMs. Also, sampling from the latter epochs is advantage in computation cost. We also analyze the curve changing of learning rate and norm of gradients in training process. The results are shown in Figure 4. According to the deduction in Section 3.3, to better approximate the posterior, the learning rate needs to be small while the norm of gradients needs to be larger than zero. Hence, the sequence needs to be sampled from about epoch 80 to 140 for this training stage.

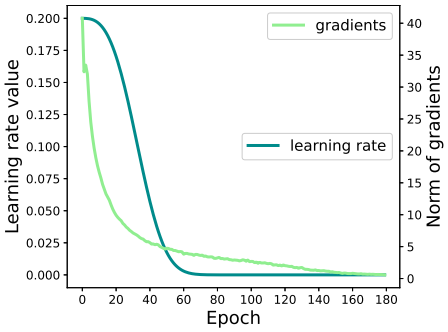


Fig. 4: Value of learning rate and the Frobenius norm of gradients in the training procedure on CIFAR100 with 5000 labels. The gradients are obtained from the last fully-connected layer

Transferring to other network architecture. We also verify the proposed approach by using WRN-28-8 as the backbone for semi-supervised learning. The experiments are based on Mean Teacher [48] and FixMatch [47] and are conducted on CIFAR100 and Mini-ImageNet with 10000 labels. For MT, it achieves 24.99% and 34.06% test error respectively, which outperforms the baseline by 2.93% and 3.83% gains. For FixMatch, the error rates are 22.58% and 21.11% for two datasets, and brings 0.75% and 0.94% than the baseline. As a comparison, MixMatch [4] and UDA [53] achieves 28.31% and 24.50% test error on CIFAR100 in the same setting respectively. These results are inferior to our approach, which demonstrates the effectiveness of our approach by a stronger backbone.

Robustness to Hyperparameters. Our approach introduces two new hyperparameters, namely the balance coefficient μ and the fused coefficient α. Here we investigate the effect of different choice for them. The parameter μ plays the role of balancing the high frequency part and the direct component. As shown in the subfigure (c) in Figure 3, we can observe that $\mu = 0.1$ achieves the best results on both two datasets. When μ becomes too large, the performance degrades obviously. It shows the significance of making a balance between the vibration baseline and its intensity. We set $\mu = 0.1$ in all our experiments for the robustness of our approach to it. The parameter α is used to balance the two components in the fused measure. As shown in the subfigure (b) in Figure 3, setting α to 0.6 and 0.2 achieves the best accuracy respectively on CIFAR10 and CIFAR100. Also, the performance for different α varies relatively smoothly on two datasets. In general, the proposed uncertainty estimation approach enjoys flexibility in hyperparameter adjustment for its robustness and limited numbers.

5 Conclusions

In this paper, we propose a novel approach for uncertainty estimation, and use it to improve learning from limited supervision. The conventional methods including the probabilities and the entropy often estimate uncertainty from instantaneous information. Different from them, we do this by using the sequential data from training process, *e.g.*, the probabilities. In particular, we measure the vibration of the obtained sequence via Fourier Transformation. By equipping with label flipping, a more accurate estimation will be obtained. Inspired by the Bayesian theory which provides a probabilistic representation of uncertainty, the sequence is sampled from latter optimization iterations. The effectiveness of the proposed approach for semi-supervised learning, active learning and one-bit supervision is verified by the extensive experiments on CIFAR10, CIFAR100, Mini-ImageNet and ImageNet.

Acknowledgements This work was supported by the National Key Research and Development Program of China under grant 2019YFA0706200, 2018AAA0102002, and in part by the National Natural Science Foundation of China under grant 61732007, 61932009.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
2. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019)
3. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019)
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
5. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning. pp. 1613–1622. PMLR (2015)
6. Cascante-Bonilla, P., Tan, F., Qi, Y., Ordonez, V.: Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. arXiv preprint arXiv:2001.06001 (2020)
7. Chen, T., Fox, E., Guestrin, C.: Stochastic gradient hamiltonian monte carlo. In: International conference on machine learning. pp. 1683–1691. PMLR (2014)
8. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
9. Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R.D., Neven, H.: Bayesian sampling using stochastic gradient thermostats. Advances in neural information processing systems **27** (2014)
10. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In: European conference on computer vision. pp. 562–577. Springer (2014)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
12. Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., Pfister, T.: Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In: European Conference on Computer Vision. pp. 510–526. Springer (2020)
13. Gastaldi, X.: Shake-shake regularization. arXiv preprint arXiv:1705.07485 (2017)
14. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. arXiv preprint arXiv:1810.12890 (2018)
15. Grandvalet, Y., Bengio, Y., et al.: Semi-supervised learning by entropy minimization. CAP **367**, 281–296 (2005)
16. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
17. Han, T., Tu, W.W., Li, Y.F.: Explanation consistency training: Facilitating consistency-based semi-supervised learning with interpretability. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7639–7646 (2021)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hu, H., Xie, L., Du, Z., Hong, R., Tian, Q.: One-bit supervision for image classification. Advances in Neural Information Processing Systems **33** (2020)
20. Hu, Z., Yang, Z., Hu, X., Nevatia, R.: Simple: Similar pseudo label exploitation for semi-supervised classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15099–15108 (2021)
21. Huang, S., Wang, T., Xiong, H., Huan, J., Dou, D.: Semi-supervised active learning with temporal output discrepancy. arXiv preprint arXiv:2107.14153 (2021)
22. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5070–5079 (2019)
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
24. Kong, L., Sun, J., Zhang, C.: Sde-net: Equipping deep neural networks with uncertainty estimates. arXiv preprint arXiv:2008.10546 (2020)
25. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
26. Kuo, C.W., Ma, C.Y., Huang, J.B., Kira, Z.: Featmatch: Feature-based augmentation for semi-supervised learning. In: European Conference on Computer Vision. pp. 479–495. Springer (2020)
27. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
28. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
29. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
30. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR'94. pp. 3–12. Springer (1994)
31. Louizos, C., Welling, M.: Multiplicative normalizing flows for variational bayesian neural networks. In: International Conference on Machine Learning. pp. 2218–2227. PMLR (2017)
32. Luo, W., Schwing, A., Urtasun, R.: Latent structured active learning. Advances in Neural Information Processing Systems **26**, 728–736 (2013)
33. Mandt, S., Hoffman, M.D., Blei, D.M.: Stochastic gradient descent as approximate bayesian inference. arXiv preprint arXiv:1704.04289 (2017)
34. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: 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)
35. Nassar, I., Herath, S., Abbasnejad, E., Buntine, W., Haffari, G.: All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7241–7250 (2021)
36. Neal, R.M.: Bayesian learning for neural networks, vol. 118. Springer Science & Business Media (2012)
37. Paisley, J., Blei, D., Jordan, M.: Variational bayesian inference with stochastic search. arXiv preprint arXiv:1206.6430 (2012)

38. Pinsler, R., Gordon, J., Nalisnick, E., Hernández-Lobato, J.M.: Bayesian batch active learning as sparse subset approximation. Advances in neural information processing systems **32**, 6359–6370 (2019)
39. Pleiss, G., Zhang, T., Elenberg, E.R., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking. arXiv preprint arXiv:2001.10528 (2020)
40. Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T.: Semi-supervised learning with ladder networks. arXiv preprint arXiv:1507.02672 (2015)
41. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)
42. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: International Conference on Learning Representations (2020)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
44. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
45. Shi, W., Yu, Q.: Integrating bayesian and discriminative sparse kernel machines for multi-class active learning. Advances in neural information processing systems (2019)
46. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5972–5981 (2019)
47. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
48. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
49. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 648–656 (2015)
50. Toneva, M., Sordoni, A., Combes, R.T.d., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. arXiv preprint arXiv:1812.05159 (2018)
51. Uhlenbeck, G.E., Ornstein, L.S.: On the theory of the brownian motion. Physical review **36**(5), 823 (1930)
52. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology **27**(12), 2591–2600 (2016)
53. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
54. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 93–102 (2019)
55. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
56. Zhang, L., Qi, G.J.: 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)