MixBag: Bag-Level Data Augmentation for Learning from Label Proportions

Takanori Asanomi[1], Shinnosuke Matsuo[1], Daiki Suehiro[1,2], Ryoma Bise[1]
[1]Kyushu University, Fukuoka, Japan [2]RIKEN AIP, Japan
{takanori.asanomi@human., shinnosuke.matsuo@human., suehiro@, bise@}ait.kyushu-u.ac.jp

Abstract

Learning from label proportions (LLP) is a promising weakly supervised learning problem. In LLP, a set of instances (bag) has label proportions, but no instance-level labels are given. LLP aims to train an instance-level classifier by using the label proportions of the bag. In this paper, we propose a bag-level data augmentation method for LLP called MixBag, based on the key observation from our preliminary experiments; that the instance-level classification accuracy improves as the number of labeled bags increases even though the total number of instances is fixed. We also propose a confidence interval loss designed based on statistical theory to use the augmented bags effectively. To the best of our knowledge, this is the first attempt to propose bag-level data augmentation for LLP. The advantage of MixBag is that it can be applied to instance-level data augmentation techniques and any LLP method that uses the proportion loss. Experimental results demonstrate this advantage and the effectiveness of our method.

1. Introduction

In general classification tasks, a model is trained on datasets where each piece of data (instance) has class labels. However, instance-level annotations cost a lot, and there are many cases in which instance-level annotation cannot be disclosed to the public for privacy reasons [23]. In such a situation, we only know the label proportions given to a set of instances (bag) and must train a model from these label proportions to classify each instance. This problem setting is known as learning from label proportions (LLP) [1].

Figure 1 illustrates the problem setup of LLP. A bag B consists of instances (e.g., images). Bag-level labels are given as training data instead of instance-level labels (class labels of each instance $\mathbf{x}$ are unknown). The label proportion $\mathbf{p}^i$ of bag B^i is given as a bag-level label. For example, if a bag B^0 contains 60, 100, and 40 instances of class 1, 2, and 3, respectively, $\mathbf{p}^i = (0.3, 0.5, 0.2)^{\mathrm{T}}$. LLP aims to train an instance-level classifier by using the label proportions of a bag. It is a weakly supervised learning task.

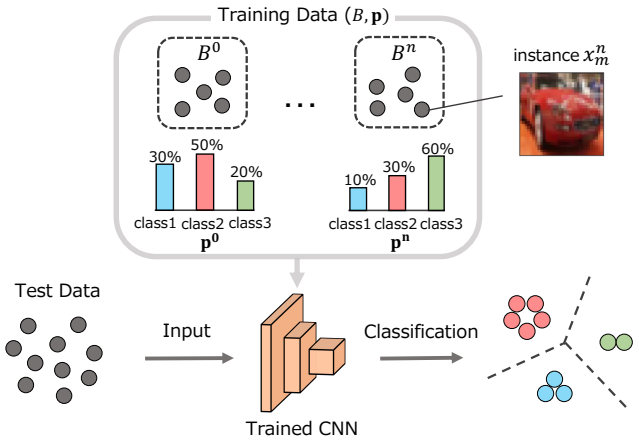

Figure 1. Illustration of learning from label proportions. LLP aims to train an instance-level classifier by using label proportions of a bag as training data instead of instance-level labels.

Many LLP methods have been proposed to address this challenging task [1, 24, 28, 27, 18, 19]. Most of them are based on the *proportion loss*, which evaluates the difference between the given proportion $\mathbf{p}^i$ and the estimated proportion $\hat{\mathbf{p}}^i$, and can be calculated by taking the average of the class probabilities of the instances in a bag. The methods based on the proportion loss work well when bag-level labels (label proportions) are sufficient. However, the accuracy decreases as the number of labeled bags decreases, i.e., insufficient labeled data causes difficulty in training a network, which is known in machine learning tasks.

In such cases, instance-level data augmentation is often an effective way to improve accuracy. Many instance-level data augmentation techniques, such as perturbation, for general classification tasks have been proposed, and their effectiveness has been demonstrated [25]. Instance-level data augmentation may improve the accuracy in LLP. However, it does not increase the number of labeled bags from the original one, and it is difficult to generate various bags with different proportions without using instance-level labels (these are unknown in LLP).

This paper proposes a *bag-level* data augmentation for

LLP that can generate new bags with various label proportions. To design an effective *bag-level* data augmentation technique, we conducted a detailed analysis of what situation can improve the accuracy in LLP. As a result, we found an important observation that the accuracy improves as the number of labeled bags increases, even though the total number of instances is fixed, i.e., different bags overlap each other as shown in Figure 2 (Right). On the basis of this observation, we propose a *bag-level* data augmentation called MixBag. MixBag increases the number of labeled bags artificially by sampling instances from a pair of original bags and mixing them: this operation mimics the above observations. The expected label proportion of a generated mixed bag can be calculated by those of the original bags. However, it may have a gap from the actual proportion of the mixed bag because randomly selected instances do not follow the proportion of the original ones. This gap adversely affects the training.

We thus introduce a confidence interval loss that helps to train a classification network by using the generated bags while avoiding this adverse effect by a proportion gap; it is statistically guaranteed. To the best of our knowledge, this is the first attempt to propose *bag-level* data augmentation for LLP. Experiments using eight datasets demonstrate the effectiveness of our method in various cases; our method improved the classification accuracy on all datasets. Our main contributions are summarized as follows.

- We examined how the number of labeled bags and the bag size affect the performance in LLP. From the preliminary experiments, we concluded that the accuracy improves as the number of labeled bags increases, even though the total number of instances is fixed.
- We proposed MixBag with a confidence interval loss. This method artificially increases the number of labeled bags, and the confidence interval loss helps train a classification network using the generated bags.
- We demonstrated that MixBag can be applied to any of the current LLP methods and any of the instance-level augmentation methods. Experimental results show that our method improved classification accuracy on various datasets.

2. Related work

2.1. Learning from Label Proportion (LLP)

Proportion loss [1] is the most common approach to LLP. To compute the proportion loss, the average of the output probability in a bag is calculated as a proportion estimation, and then the bag-level cross-entropy between the true label proportion and the estimated proportion is used as the loss.

The proportion loss has been extended in many methods; e.g., introducing regularization terms [33, 1, 8, 28, 30, 17, 18]. Tsai *et al.* [28] incorporated a consistency regularization into the proportion loss, which forces the network predictions to be consistent when its input is perturbed. Yang *et al.* [30] use contrastive learning for pre-training in a self-supervised learning manner and then train the model by using the proportion loss. LLP-GAN [17] introduces the generative adversarial network (GAN) to LLP. The fake instances created by the generator encourage the discriminator to distinguish real or fake and classify the real instances. This method also uses the proportion loss to train the classifier. Two-stage LLP [18] introduces pseudo-labeling after pre-training by using a proportion loss.

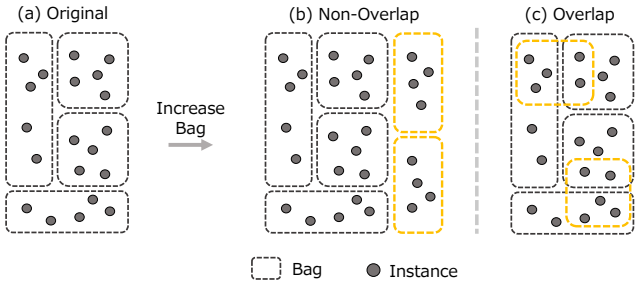

Figure 2. Illustration of overlap and non-overlap situations. (a) Original bags. (b) New bags (yellow) are added with non-overlap, i.e., the total number of instances increases. (c) New bags overlap with the original ones, i.e., the total number of instances is fixed.

It is known that the classification accuracy decreases when the labeled data is insufficient in machine learning. To deal with this issue, data augmentation is a standard approach. However, *bag-level* data augmentation for LLP is still an unexplored area. This paper proposes a simple but effective *bag-level* data augmentation method. One of the advantages of our method is that any LLP method can be applied after our *bag-level* augmentation.

2.2. Instance-level data Augmentation

Data augmentation is a standard way to address the problem of inadequately labeled data. Many instance-level data augmentation techniques have been proposed. Basic data augmentation involves adding perturbations to the original images, such as by flipping the image [25], adding noise [25], and erasing parts of an image [7, 2, 14]. Image mix [11, 35, 34, 10] is another approach, where methods such as MixUP [35]and CutMix [34] mix together two original images. GAN-based augmentation methods [22, 36, 26] represent data distributions and generate new images based on these distributions. Auto augment [3, 16, 4] automatically searches for suitable augmentation approaches that improved performance. The above methods are instance-level data augmentations. It is not a simple task to generate various bags that have different proportions without using instance-level labels. For example, CutMix and MixUp require the labels of the original im-

ages (instances) in order to prepare the label proportions of a generated new bag. Since simple perturbation methods do not require the ground truth of the image label for augmentation, they can be used to add perturbations to each instance in a bag. However, these methods are not designed for *bag-level* data augmentation.

2.3. Bag-Level data augmentation for MIL

Few *bag-level* data augmentation methods have been proposed for multiple instance learning (MIL), which trains a classifier from bag-level class labels (positive or negative). Li et al. [15] proposed a MIL-based method that uses *bag-level* augmentation for COVID-19 severity assessment. This method takes a pseudo-labeling approach, where positive and negative instances are taken from positive bags based on the estimated confidence of the instance class and used to make a new bag. Yang et al. [31] proposed a MIL method for segmenting pathology images. This method generates new bags by clustering instances. It is reasonable for MIL to make a positive or negative bag from high confidence positive or negative instances. However, it is difficult to make a proportion label from pseudo labeling since all of the pseudo labels have to be accurate to calculate the correct proportion. So in order to improve the performance of LLP, appropriately designing a *bag-level* data augmentation is required.

3. Preliminary: LLP and Proportion Loss

In LLP, the training set contains n bags, $B^1, \dots, B^n$. Each bag B^i consists of a set of instances, i.e., $B^i = \{x^i_j\}_{j=1}^{|B^i|}$, and a vector of label proportions $\mathbf{p}^i = (p^i_1, \dots, p^i_c, \dots, p^i_C)^T$ is attached to each bag as training label, where C is the number of classes. This label indicates that $p^i_c \times |B^i|$ instances in the i-th bag belong to the class c for any $c \in \{1, \dots, C\}$. The goal of LLP is to train a network that estimates the class label of each instance by using only the proportion labels of the bags. Proportion loss [1] is a standard approach to this challenging task, and many methods are based on it.

Given training data with label proportions, the standard proportion loss is a bag-level cross-entropy between the ground-truth of the label proportion $\mathbf{p}^i$ and the predicted proportion $\hat{\mathbf{p}}^i$, which is the average of the output class probabilities in a bag. It is defined as:

$$\ell_{\mathrm{prop}}(B^i, \mathbf{p}^i, f) = -\sum_{c=1}^{C} p^i_c \log \hat{p}^i_c, \tag{1}$$

$$\hat{p}^i_c = \frac{1}{|B^i|} \sum_{j=1}^{|B^i|} f(x^i_j)_c, \tag{2}$$

where $f(x^i_j)_c$ is the output probability that x^i_j belongs to c.

4. Preliminary Empirical Analysis of LLP

In previous works, the relationship between the number of labeled bags, bag sizes, and accuracy in LLP has not been analyzed well. Some papers [17, 8] reported that the accuracy increased as the larger bag size in their experiments. However, the reason was not discussed well. In their experiments, they changed the bag size while the total number of instances was fixed. Since an instance generally does not belong to two bags (i.e., bags does not overlap each other) in LLP, it indicates that the bag size and the number of labeled bags have a trade-off; when the bag size increases, the number of labeled bags decreases.

In this section, we further empirically analyze the relationships among the number of labeled bags, bag sizes, and accuracy by using the proportion loss on various public datasets. First, we conducted experiments on whether the number of labeled bags or the bag size has more influence on the classification accuracy in a general setup of LLP, where different bags do not contain the same instance (no overlap), as shown in Figure 2 (b). To avoid the trade-off between the number of labeled bags and the bag size, we fixed the number of labeled bags to analyze the effect for accuracy by the number of labeled bags, and vice versa. In the experiments, the total number of instances changed. We speculated that this analysis may give clues for developing a new method in LLP.

Datasets and experimental settings: We used eight datasets in our experiments[1]: 1) CIFAR10 [12], which is commonly used in classification tasks. The dataset contains ten classes, such as vehicles and animals; 2) SVHN [20], which is a house-number-images dataset, and a character-level label is given like in MNIST [6]; 3) OCT, which contains four classes of Retinal OCT images; 4) BLOOD, which contains eight classes of Blood Cell Microscope images; 5) ORGANA, 6) ORGANC, 7) ORGANS, which contain eleven classes of abdominal CT images. The difference is the view of the image and its sizes, and 8) PATH, which contains nine classes of colon pathology images. Note that 3) to 8) are datasets in MedMNIST [32], and we did not use the small datasets in [32] since they cannot be used to prepare enough bags for LLP.

To prepare a bag, the proportions of each class were randomly selected from a Gaussian distribution; the decided proportion was used as the label of the bag. Then, samples were randomly selected from each class, in which the number of samples for each class followed the proportion of the class. In real applications, different bags do not contain the same instance; i.e., they do not overlap. We thus generated bags without overlap in these experiments.

Performance when changing the number of labeled bags

[1] Almost all papers on LLP use CIFAR10 and SVHN

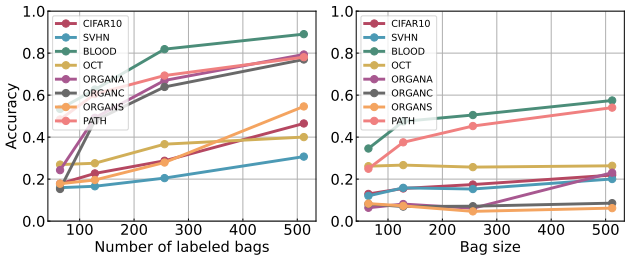

Figure 3. Relationship among the number of labeled bags, bag sizes, and accuracy in a general setup of LLP, where different bags do not contain the same instance (no overlap). **Left**: Accuracy improvement when increasing the number of labeled bags (bag size is fixed). **Right**: Accuracy improvement when increasing the bag size (the number of labeled bags is fixed).

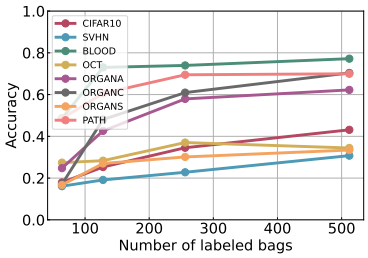

Figure 4. Experimental results on how the number of labeled bags affects the accuracy in the case of overlap, i.e., when the total number of instances was fixed.

in the general setup of LLP (no overlap): We examined how the number of labeled bags affects the instance-level classification accuracy for the standard proportion loss. We varied the number of labeled bags (64, 128, 256, 512) while keeping the bag size fixed at 10. In this experiment, we follow the general setup of LLP; different bags do not overlap, i.e., the number of instances changed when the number of labeled bags was changed. As shown in Figure3 (Left), the accuracy improved due to increasing the number of labeled bags in all datasets. This is reasonable because the number of labeled bags indicates the number of proportion labels; more labeled data for training is expected to improve accuracy in machine learning.

Performance when changing the bag size in the general setup of LLP (no overlap): We analyzed how the bag size affects the accuracy in the general setup of LLP. In this experiment, we changed the bag size (64, 128, 256, 512) but fixed the number of labeled bags at 10. As shown in Figure 3 (Right), the accuracy did not vary significantly with the bag size in six of the datasets, while it improved in two datasets. This result shows that increasing the number of instances without increasing the labels does not effectively improve accuracy.

Performance when changing the number of labeled bags in the case of overlap: The above observations indicate that increasing the number of labeled bags is important to improving accuracy. Another factor that might have improved the accuracy is increasing the total number of 'instances,' which occurs when the number of labeled bags increases. To confirm whether the number of labeled bags or the total number of instances affects accuracy, we performed experiments on how the number of labeled bags affects the accuracy in the case of overlap, i.e., when the total number of instances was fixed. The initial bags were prepared, where the number of labeled bags was 64, and the bag size was 10. To increase the number of labeled bags, we then incrementally added new bags (64, 128, 256), where the new bags were generated from the initial instances that belong to the initial bags, i.e., generated bags overlap with the initial ones, as shown in Figure 2 (c). As shown in Figure4, the accuracy improves as the number of labeled bags increases in all datasets. This observation is important for our *bag-level* data augmentation. If we were to mimic this situation where the number of labeled bags increases and the total number of instances is fixed, it should improve accuracy, even though a situation with bags having overlaps is unnatural in actual applications. However, we can make this situation by generating new bags from the original ones.

5. MixBag with confidence interval loss

To mimic the situation of increasing the number of labeled bags and allowing their overlap, we propose MixBag, which is a *bag-level* data augmentation technique.

5.1. MixBag

Figure 5 (Left) shows an overview of MixBag. Given two randomly selected bags $\{B^i, B^j\}$ where each bag is labeled with a proportion, $\mathbf{p}_i$ or $\mathbf{p}_j$, mix the two bags to generate a new bag. More precisely, $n_i = |B^i| \times \gamma$ and $n_j = |B^j| \times (1-\gamma)$ instances are randomly sampled from B^i and B^j, respectively, where $\gamma \in [0, 1]$ is a random variable, and the sampled sub-bags are denoted as S^i and S^j. Then, these two sub-bags are combined to generate a mixed bag, $B^k = S^i \cup S^j$. This generated mixed bag overlaps with the original bags while the total number of instances is fixed, and the number of labeled bags is increased by the generated bags. It can be considered to mimic the third preliminary experiment (Figure 2 (c)).

5.2. Confidence interval loss

The sampled sub-bags S^i, S^j are expected to have almost the same proportions as the original ones, $\mathbf{p}^i$ and $\mathbf{p}^j$, if the randomly selected instances follow the proportion of the original bags. In this case, the expected proportion of the generated new bag B^k is $\mathbf{p}^k = \gamma\mathbf{p}^i + (1-\gamma)\mathbf{p}^j$. However, in practice, the actual proportion of randomly sampled instances may have some gaps from that of the original bag

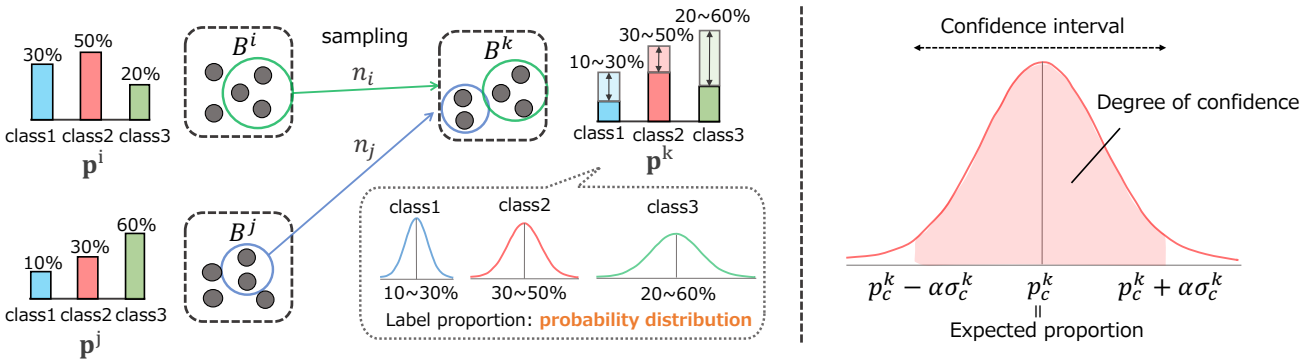

Figure 5. Overview of our method. **Left**: the way to create mixed bags and illustration of label proportion's probability. **Right**: confidence interval

because the number of sampled instances in each class does not always follow the original proportion. This gap can be estimated by calculating the confidence interval (CI) for a population proportion.

Let us denote the proportion of a class c in the i-th bag as p_c^i. The standard deviation of the class c proportion of the sampled sub-bag S^i is defined as:

$$\sigma_c^i = \sqrt{\frac{p_c^i(1-p_c^i)}{n_i}}, \tag{3}$$

where $n_i = |B^i| \times \gamma$ is the number of instances randomly sampled from B^i. This equation is well-known in statistics; it can be derived using the central limit theorem.

The CI for the truth proportion $\tilde{p}_c^i$ of the sub-bag S^i is formulated as:

$$p_c^i - \alpha\sigma_c^i \leq \tilde{p}_c^i \leq p_c^i + \alpha\sigma_c^i, \tag{4}$$

where α is set according to the desired degree of confidence, e.g., 95%. If we set the confidence to a higher value, α becomes higher values. For the other sub-bag S_j, the CI for the population proportion of $\tilde{p}_c^j$ is formulated in the same manner.

The CI for $\tilde{p}_c^k$, which is the proportion of the newly generated bag B^k, is defined as:

$$p_c^k - \alpha\sigma_c^k \leq \tilde{p}_c^k \leq p_c^k + \alpha\sigma_c^k, \tag{5}$$

$$p_c^k = \gamma p_c^i + (1-\gamma)p_c^j, \tag{6}$$

$$\gamma = \frac{n_i}{n_i + n_j}, \tag{7}$$

$$\sigma_c^k = \gamma\sqrt{\frac{p_c^i(1-p_c^i)}{n_i}} + (1-\gamma)\sqrt{\frac{p_c^j(1-p_c^j)}{n_j}}, \tag{8}$$

where p_c^k is the expected proportion of the generated mixed bag, which is directly computed from the proportions of the original bags B^i, B^j, and the standard deviation σ_c^k of the mixed bag can be calculated by using the standard deviation of the sub-bags with the ratio γ.

This confidence interval (Eq.5) indicates that the actual proportion of the mixed bag has a gap from the expected proportion. If we directly use the proportion $\mathbf{p}^k$ (Eq.6) for the mixed bag as augmented data and use it for the proportion loss, the gap from the actual proportion may adversely affect the training of the network since the gap causes a label noise.

To avoid this adverse effect caused by gaps (label noise), we propose a confidence interval loss (CI loss) for generated mixed bags. The CI loss is based on the proportion loss [1]. In contrast to the standard proportion loss, the CI loss ignores a loss when the estimated proportion $\hat{p}_c^k$ is in the confidence interval $[p_c^k - \alpha\sigma_c^k, p_c^k + \alpha\sigma_c^k]$ since gaps in the CI often occur.

Given the degree of confidence $\alpha\%$ (the corresponding value is set to α), such as when the degree of confidence is 95%, $\alpha = 1.96$, the CI loss is defined as:

$$\ell_{\mathrm{CI}}(B^k, \mathbf{p}^k, f) = -\sum_{c=1}^{C} \mathbf{1}(p_c^k, \hat{p}_c^k) p_c^k \log \hat{p}_c^k, \tag{9}$$

$$\hat{p}_c^k = \frac{1}{|B^k|}\sum_{j=1}^{|B^k|} f(x_j^k)_c, \tag{10}$$

$$\mathbf{1}(p, p') = \begin{cases} 0 & \text{if } p - \alpha\sigma \leq p' \leq p + \alpha\sigma \\ 1 & \text{otherwise} \end{cases}, \tag{11}$$

where $\mathbf{1}$ is the indicator function, which outputs 0 if p' is in the confidence interval, and 1 otherwise. This CI loss updates the network parameters if the difference between the estimated proportion $\hat{p}_c^k$ and the expected proportion p_c^k calculated by Eq.6 is significant. Using this loss, we can accelerate the training by using the augmented bags without any adverse effects from proportion gaps.

In training, mixed bags are randomly generated in each batch. It indicates that the number of generated bags will increase with iteration. We stop training based on the propor-

Method	CIFAR10 Accuracy	SVHN Accuracy	PATH Accuracy	OCT Accuracy	BLOOD Accuracy	ORGANA Accuracy	ORGANC Accuracy	ORGANS Accuracy	Average Accuracy
LLP [1]	0.4538	0.3009	0.7843	0.4336	0.8869	0.7635	0.7898	0.5372	0.6187
LLP + Ours(w/o CI)	0.4582	0.2971	0.7884	0.425	0.8898	0.7831	0.8189	0.563	0.6280
LLP + Ours	0.5256	0.3742	0.7861	0.4347	0.9017	0.7971	0.8099	0.6197	**0.6561**
LLP + Ours(supervised)	0.6594	0.7781	0.8118	0.5396	0.8947	0.8222	0.8233	0.6674	0.7234
LLP-VAT [28]	0.4740	0.3119	0.8915	0.4124	0.8915	0.7936	0.8060	0.5837	0.6455
LLP-VAT + Ours(w/o CI)	0.5203	0.4302	0.8779	0.4503	0.8779	0.7847	0.7656	0.5849	0.6614
LLP-VAT + Ours	0.5283	0.4111	0.8944	0.4366	0.8944	0.8155	0.8228	0.6376	**0.6800**
LLP-VAT + Ours(supervised)	0.5963	0.5976	0.8997	0.4777	0.8997	0.8230	0.8313	0.6622	0.7234
LLP-PI [13]	0.4702	0.3011	0.8886	0.4224	0.8862	0.7810	0.7962	0.5917	0.6421
LLP-PI + Ours(w/o CI)	0.4794	0.5209	0.8688	0.4350	0.8573	0.7595	0.7570	0.5741	0.6565
LLP-PI + Ours	0.5290	0.3865	0.8841	0.4366	0.8841	0.8013	0.7989	0.6236	**0.6680**
LLP-PI + Ours(supervised)	0.5896	0.7721	0.8918	0.4720	0.8918	0.8101	0.8184	0.6577	0.7379

Table 1. Instance-level classification accuracy when our method was applied to three baseline methods; LLP [1], LLP-VAT [28] and LLP-PI [13]. Each value is the average of five-fold cross-validation results. Here, 'CI' means the confident interval, and 'supervised' indicates the case when a ground-truth proportion to the generated mixed bag is given.

Method	CIFAR10 Accuracy	SVHN Accuracy	PATH Accuracy	OCT Accuracy	BLOOD Accuracy	ORGANA Accuracy	ORGANC Accuracy	ORGANS Accuracy	Average Accuracy
LLP [1]	0.4538	0.3009	0.7843	0.4336	0.8869	0.7635	0.7898	0.5372	0.6187
LLP + Ours	0.5256	0.3742	0.7861	0.4347	0.9017	0.7971	0.8099	0.6197	**0.6561**
LLP + Flip	0.5777	0.3884	0.8299	0.4793	0.9265	0.6527	0.6904	0.6293	0.6467
LLP + Flip + Ours	0.5764	0.4124	0.8005	0.5504	0.9155	0.6718	0.6717	0.6280	**0.6533**
LLP + Erase	0.5551	0.3500	0.8097	0.4414	0.9094	0.8393	0.8447	0.6578	0.6759
LLP + Erase + Ours	0.5671	0.4269	0.7934	0.4291	0.9021	0.8266	0.8313	0.6594	**0.6794**
LLP + Invert	0.5258	0.6763	0.7775	0.4394	0.8790	0.7779	0.8215	0.6255	0.6899
LLP + Invert + Ours	0.5784	0.7761	0.7421	0.4580	0.8736	0.7845	0.8323	0.6329	**0.7097**
LLP + Gaussianblur	0.4766	0.3686	0.7639	0.4145	0.8875	0.7906	0.8318	0.6325	0.6457
LLP + Gaussianblur + Ours	0.5135	0.4036	0.7555	0.4300	0.8789	0.7962	0.7989	0.6709	**0.6559**
LLP + Persperctive	0.5675	0.5441	0.7892	0.4678	0.8968	0.8309	0.8440	0.6689	0.7011
LLP + Perspective + Ours	0.5750	0.5715	0.7745	0.4962	0.9064	0.8180	0.8404	0.6895	**0.7089**

Table 2. Instance-level classification accuracy when our method was applied after various instance-level augmentation methods.

tion loss in validation data. Note that instance-level labels cannot be used for validation in LLP.

6. Experiments

6.1. Datasets and experimental settings

Preparation of bags: The eight datasets used in these experiments were the same as those used in the preliminary experiments. The way of generating bags was also the same as in the preliminary experiments. Although some of the related studies generated bags by allowing overlaps to increase the number of labeled bags, we did not allow overlaps since this situation is unnatural in actual applications. To keep the experiment setups the same in all datasets, we set the number of labeled bags as 512 and the bag size as 10 based on the smallest dataset. This small bag size can be considered a challenging scenario for MixBag since the number of samplings for sub-bags decreases, and it increases the confidence interval. We also evaluated our method using different bag sizes to demonstrate the effectiveness of our method in various situations.

Implementation: We implemented our method by using Pytorch [21]. The network model was ResNet18 [9] pre-trained on the ImageNet dataset [5]. To train our network, we also used the Adam Optimizer [9] with a learning rate of 3e-4, epoch$= 1000$, mini-batch size $= 32$, early stopping $= 10$ and trained on an NVIDIA GeForce 3090 GPU. We set the confidence degree as '99%' and selected γ randomly from a uniform distribution $[0, 1]$.

Evaluation metric: We evaluated the model performance by the instance-level classification accuracies, which were calculated by using the confusion matrix. To compute the metric, we performed a five-fold cross-validation in all our experiments. The accuracies listed in the tables are the average values of the five-fold cross-validation.

6.2. Comparison

Comparison with current methods: Our method MixBag can be applied to any proportion loss-based method since MixBag is a data augmentation method, in which any method can be applied after data augmentation with the CI loss. Most current LLP methods incorporate self-supervised learning, such as consistency, into the proportion loss. We

Method	γ-sampling	Average Accuracy
LLP [1]	–	0.6187
LLP + Ours	uniform	**0.6561**
LLP + Ours	Gauss	0.6512
LLP + Ours	half	0.6524

Table 3. Average accuracy on eight datasets when changing a sampling method for γ.

Method	CI	Average Accuracy
LLP [1]		0.6187
LLP + Ours(99%)	✓	**0.6561**
LLP + Ours(95%)	✓	0.6519
LLP + Ours(80%)	✓	0.6531
LLP + Ours(50%)	✓	0.6243

Table 4. Average accuracy on eight datasets when changing the degree of confidence interval (50%, 80%, 95%, 99%).

thus applied our method in three methods: 1) LLP [1], which is the method of the original paper on using the proportion loss; 2) LLP-VAT [28], which introduces the consistency regularization into the proportion loss; and 3) LLP-PI [13], which uses the additive Gaussian noise for making perturbed examples. The LLP-VAT is one of the state-of-the-art LLP methods using the proportion loss, and its implementations are publicly available.

We compared the baseline methods as an ablation study. The first was 'LLP', which uses the standard proportion loss; the second was 'LLP+Ours(w/o CI),' which uses the MixBag without using the CI loss, where the proportion is directly calculated from the original bags (and may contain noise) was used as the label for the generated mixed bag. The third was 'LLP+Ours(supervised)', which uses the ground-truth proportions of the mixed bags. Note that the proportion would be unknown in real cases.

Table 1 shows the accuracy of each method, where the gray color field indicates the proposed methods in each baseline method. 'LLP+Ours(w/o CI)' improved the accuracies in some datasets and slightly improved them on average from the baselines. However, the improvement was limited by the noisy proportions. MixBag improved the accuracy of every baseline method on all datasets, and the improvement was 3 to 4 points on average.

Introducing MixBag with instance-level augmentation: Next, we introduced our method to the instance-level augmentation methods. We used standard augmentation techniques: 'Flip' randomly flips the image horizontally and upside down. 'Erase' removes a randomly selected area from an original image. 'Invert' randomly changes the colors of the image by inversion. 'Gaussianblur' blurs an image by adding a randomly chosen Gaussian blur. 'Perspective' performs a random projective transformation on an image. To generate a bag by instance-level augmentation, we added perturbations into instances randomly selected from an original bag; i.e., an augmented bag contains the perturbated instances and original instances, where this augmentation has not been reported in previous papers. In LLP + instance-level + MixBag, the mixed bag was created from the original and instance-level augmented bags. Table 2 shows the accuracies of the baseline and our methods. The instance-level data augmentation improved the performance from the baseline. Our method further improved the accuracies of every instance-level augmentation method.

Method	Size	Num	Average Accuracy
LLP [1]	10	512	0.6187
LLP + Ours(w/o CI)	10	512	0.6280
LLP + Ours	10	512	**0.6561**
LLP [1]	20	256	0.5441
LLP + Ours(w/o CI)	20	256	0.5496
LLP + Ours	20	256	**0.5667**
LLP [1]	40	128	0.4746
LLP + Ours(w/o CI)	40	128	0.4600
LLP + Ours	40	128	**0.5058**

Table 5. Average accuracy on eight datasets when changing the number of labeled bags and bag size. 'Size' means the bag size. 'Num' means the number of labeled bags.

6.3. Performance in various situations

Performance with different γ sampling methods: We evaluated our method with three different γ sampling methods, i.e., three different ways to determine the mixing rate γ of the two sub-bags S^i, S^j: 'uniform' selects γ randomly from a uniform distribution; 'Gauss' selects γ randomly from a Gaussian distribution $\mathcal{N}(0.5, 0.25)$; 'half' always set γ to 0.5. Table 3 shows the average accuracies of all datasets. All sampling methods outperformed the performance compared to the baseline method (LLP); 'uniform' was marginally better than other methods.

Performance with different confidence intervals: We conducted experiments with four different confidence intervals, 99%, 95%, 80%, and 50%. Table 4 shows the average accuracies of all datasets. For all confidence interval settings, our method improved the accuracy on all datasets and in all cases. When the confidence interval took high values (99%, 95%, 80%), our method worked well since the worst effects of noisy labels were mitigated by the CI loss. Even when the confidence interval was low (50%), the accuracy improved from the baseline method. The case of 99% was the best on average. We consider that a larger confidence interval avoids the adverse effects of noisy proportions and makes the training more stable.

Performance when changing the number of labeled bags and bag size: To show the effectiveness of our method in various situations, we conducted experiments with different bag sizes (512, 256, 128) and different numbers of labeled bags (10, 20, 40), where the bag size and the number of

Method	Bag-Generation	CI	Average Accuracy
LLP [1]	–	–	0.6187
LLP + Ours(w/o CI)	Union	–	0.6226
LLP + Ours(w/o CI)	Sub-bag		0.4956
LLP + Ours(with CI)	Sub-bag	✓	0.6364
LLP + Ours(w/o CI)	MixBag		0.6280
LLP + Ours(with CI)	MixBag	✓	**0.6561**

Table 6. Average accuracy on eight datasets in different bag-generation methods.

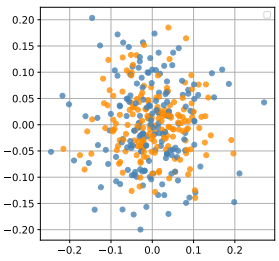

Figure 6. Distribution of proportion vectors. **Blue**: Original bag's proportion. **Orange**: Mixed bag's proportion.

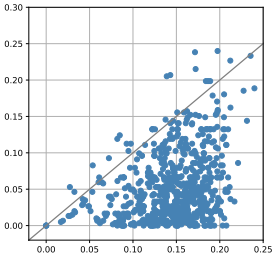

Figure 7. Relationship between confidence interval and actual gaps. **Blue**: proportion vector of a bag.

labeled bags have a trade-off since the total number of instances is fixed. Table 5 shows the average accuracies on all datasets with various setups. Our method improved the accuracy from the baseline methods on average in all cases. This demonstrates the robustness of our method.

Performance with different bag generations: To confirm the effectiveness of the 'Mix' approach, we compared MixBag with two other bag generation methods. 1) 'union' directly combines any two bags (B^i, B^j) without sampling. In this case, the proportion of a generated bag can be directly calculated from those of the original bags; i.e., there is no noise for the combined proportions by taking the union of two bags. Therefore, the CI loss was not used for this method. 2) 'sub-bag' randomly samples instances from a single bag B^i. It can be considered as a sub-bag of the original bag B^i. In this case, the label proportions of the sub-bag may have a gap from the original one, and thus, we evaluated two cases; 'sub-bag' without and with the CI loss. 3) 'MixBag' is our method, which mixes two bags. We also evaluated two cases; Mixbag with and without CI loss.

Table 6 shows the average of the instance-level accuracies of the comparative methods on the eight datasets. Since 'sub-bag with CI' and union' can increase the number of labeled bags, these methods had slightly improved accuracy from that of baseline method LLP [1]. Our method, 'MixBag with CI' further improved the accuracy compared with these other bag generation methods. We consider that our method has two characteristics of 'sub-bag' and 'union', where MixBag first makes sub-bags from original bags and then takes the union of these sub-bags. In addition, MixBag can generate various proportions of the mixed bag compared to 'sub-bag' and 'union'. Therefore, MixBag was better than them.

6.4. Detailed Analysis

Distribution of proportion vectors in the space of proportion labels: Figure 6 shows distributions of the original proportion vectors $\{\mathbf{p}^i\}_{i=1}^n$ (blue) and the generated mixed bag's proportion vectors $\{\mathbf{r}^i\}_{i=1}^m$ (orange) in the 'ORGANS' dataset, in which the dimension was compressed to 2D by PCA [29]. Here, the number of mixed bags is the same as that of the original bags. We can see that the MixBag proportion vectors do not overlap with the original ones, i.e., new proportion labels were generated, and the distribution of mixed bag proportions covers the original distribution by interpolating pairs of the original bags. This result shows that MixBag adds diversity to the training dataset, which facilitates an improvement in the instance-level classification accuracy.

Relationship between confidence interval and ground-truth proportion: Figure 7 shows how well the confidence interval works. The horizontal axis indicates the difference $(\|\mathbf{p} - \hat{\mathbf{p}}\|_1)$ between the ground-truth proportion $\mathbf{p}$ and the estimated proportion $\hat{\mathbf{p}}$ of a mixed bag. The vertical axis indicates the difference $(\alpha\boldsymbol{\sigma}^k)$ between the upper bounds of the confidence interval and $\mathbf{p}$, where we set the degree of confidence as 99%. Each blue point indicates the proportion vector of a mixed bag. Almost all points are under the straight line in the figure. This indicates that the actual proportion gaps from the ground truth were less than the confidence interval, and thus the proposed CI loss works properly.

7. Conclusion

In this paper, we examined how the number of labeled bags and the bag size affects the performance in LLP. As a result, we found that the accuracy improves as the number of labeled bags increases, even when the total number of instances is fixed. Then, based on this observation, we propose a simple but effective *bag-level* data augmentation method. In addition, we also proposed a confidence interval loss that effectively trains a classification network by using the generated bags while avoiding the adverse effects caused by noisy proportions. To the best of our knowledge, this is the first attempt to propose *bag-level* data augmentation for LLP. The experiments using eight datasets demonstrated the effectiveness of our method in various cases. Additionally, MixBag can be applied to instance-level data augmentation techniques and any LLP method that uses a standard proportion loss.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number JP20H04211 and JP23K18509, and JST, ACT-X Grant Number JPMJAX200G, Japan.

References

[1] Ehsan Mohammady Ardehaly and Aron Culotta. Co-training for demographic classification using deep learning from label proportions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1017–1024, 2017.

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020.

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[8] Dulac-Arnold Gabriel, Zeghidour Neil, Cuturi Marco, Beyer Lucas, and Vert Jean-Philippe. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909*, 2020.

[9] 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*, pages 770–778, 2016.

[10] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[11] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018.

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

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

[14] Pu Li, Xiangyang Li, and Xiang Long. Fencemask: a data augmentation approach for pre-extracted image features. *arXiv preprint arXiv:2006.07877*, 2020.

[15] Zekun Li, Wei Zhao, Feng Shi, Lei Qi, Xingzhi Xie, Ying Wei, Zhongxiang Ding, Yang Gao, Shangjie Wu, Jun Liu, et al. A novel multiple instance learning framework for covid-19 severity assessment via data augmentation and self-supervised learning. *Medical Image Analysis*, 69:101978, 2021.

[16] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Jiabin Liu, Bo Wang, Zhiquan Qi, Yingjie Tian, and Yong Shi. Learning from label proportions with generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[18] Jiabin Liu, Bo Wang, Xin Shen, Zhiquan Qi, and Yingjie Tian. Two-stage training for learning from label proportions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2737–2743, 8 2021.

[19] Shinnosuke Matsuo, Ryoma Bise, Seiichi Uchida, and Daiki Suehiro. Learning from label proportion with online pseudo-label decision by regret minimization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[20] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019.

[22] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[23] Zhiquan Qi, Bo Wang, Fan Meng, and Lingfeng Niu. Learning with label proportions via npsvm. *IEEE transactions on cybernetics*, 47(10):3293–3305, 2016.

[24] Stefan Rueping. Svm classifier estimation from group probabilities. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 911–918, 2010.

[25] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[26] Hao Tang, Hong Liu, and Nicu Sebe. Unified generative adversarial networks for controllable image-to-image translation. *IEEE Transactions on Image Processing*, 29:8916–8929, 2020.

[27] Hiroki Tokunaga, Brian Kenji Iwana, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Negative pseudo labeling using class proportion for semantic segmentation in pathology. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 430–446. Springer, 2020.

[28] Kuen-Han Tsai and Hsuan-Tien Lin. Learning from label proportions with consistency regularization. In *Asian Conference on Machine Learning*, pages 513–528. PMLR, 2020.

[29] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[30] Haoran Yang, Wanjing Zhang, and Wai Lam. A two-stage training framework with feature-label matching mechanism for learning from label proportions. In *Asian Conference on Machine Learning*, pages 1461–1476, 2021.

[31] Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pages 35–45. Springer, 2022.

[32] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[33] Shi Yong, Liu Jiabin, Wang Bo, Qi Zhiquan, and Tian YingJie. Deep learning from label proportions with labeled samples. pages 73–81, 2020.

[34] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[35] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.