Abstract

In active learning, the focus is mainly on the selection strategy of unlabeled data for enhancing the generalization capability of the next learning cycle. For this, various uncertainty measurement methods have been proposed. On the other hand, with the advent of data augmentation metrics as the regularizer on general deep learning, we notice that there can be a mutual influence between the method of unlabeled data selection and the data augmentation-based regularization techniques in active learning scenarios. Through various experiments, we confirmed that consistency-based regularization from analytical learning theory could affect the generalization capability of the classifier in combination with the existing uncertainty measurement method. By this fact, we propose a methodology to improve generalization ability, by applying data augmentation-based techniques to an active learning scenario. For the data augmentation-based regularization loss, we redefined cutout (co) and cutmix (cm) strategies as quantitative metrics and applied at both model training and unlabeled data selection steps. We have shown that the augmentation-based regularizer can lead to improved performance on the training step of active learning, while that same approach can be effectively combined with the uncertainty measurement metrics proposed so far. We used datasets such as FashionMNIST, CIFAR10, CIFAR100, and STL10 to verify the performance of the proposed active learning technique for multiple image classification tasks. Our experiments show consistent performance gains for each dataset and budget scenario. The source code will open to the public.

1. Introduction

Active learning is a method that estimates the uncertainty of unlabeled data to select candidates to be labeled which can improve the learning curve on a given budget scenario. Beyond image recognition, it has been widely studied to improve the learning efficiency in medical image recognition or semantic segmentation problems, where the cost of labels is very high [16]. Recently, as the deep learning based architectures have shown excellent performance in many fields and there is more access to libraries for deep learning, research is actively conducted to apply a deep neural network to active learning scenarios [5, 2, 15, 10, 20].

However, in active learning research using deep neural networks, the role of deep neural networks does not actively reflect the inherent characteristics of deep neural network training methods in active learning scenarios, except for the role of feature encoders that provide learning representation [5, 2, 15]. Recently, active learning research using structural features of deep neural networks has been proposed [20]. In [20], a submodule of a convolutional neural network (CNN) was used to estimate the expected loss in the process of training the learning representation for the target task, assuming uncertainty for the unlabeled data during the CNN training. However, various methodologies that utilize the characteristics of deep neural network training for active learning scenarios still need to be studied.

Furthermore, there is an unnoticed point in active learning that researches show improved performance only in specific datasets and budget scenarios. [17] showed that methods still have poor generalization performance when datasets or budget conditions change. To overcome this problem, our approach is motivated by the analytical learning theory [12]. In short, analytical learning theory suggests a way to measure the generalization gap approximately, and [12] also shows how the theory can be applied to deep learning as a consistency-based regularizer. We adopt the idea of the cutout and cutmix data augmentation techniques [4, 21] and we devised consistency-based regularization losses and uncertainty measures. Our suggesting methods pursue to improve the generalization capabilities of deep neural network models, to actively reflect the characteristics of deep neural network training in active learning. Inspired by existing research [12], we propose a learning methodology that maximizes efficiency which is insensitive to budgets and datasets by combining uncertainty measurement techniques and regularization loss for each step of active learning scenarios: candidate sampling step and training step.

Besides, various image transformation techniques are
Figure 1. A schematic depiction of the proposed active learning method. $h(L_{t-1}, t)$ is a model trained with labeled data $L_{t-1}$ in $t$ cycles in an active learning scenario, and $h(L_t, t + 1)$ is a model trained with labeled data $L_t$ in $t + 1$ cycle. $\theta(t)$ represents the decision boundary generated in the embedding space for classification from the model $h(L_{t-1}, t)$ and $\theta(t + 1)$ represents the decision boundary generated from the model $h(L_t, t + 1)$. In the embedding space, the samples indicated by the red dotted lines are the samples selected as the samples to be labeled in the $t + 1$ cycle with high uncertainties among the unlabeled samples of each class. In this case, the decision boundary generated in each cycle may be influenced by the method of sample selection and the training strategy.

applied to mini-batches for robustness during the learning process of the image recognition models [14, 18, 6]. Afterward, data augmentation-based regularization techniques [4, 21, 23], which are derived from the image transformation, have been proposed as a major factor in deep neural network training to induce direct generalization performance improvement of CNNs [7]. We attempted to reflect the characteristics of these data augmentation techniques to active learning, which are generally used in the training of deep neural network. To do this, we first observed that data augmentation-based regularization techniques could be a major factor in measuring the generalization capability of unlabeled data. In order to use this observation, we made use of analytical learning theory to estimate the uncertainty, by using the measurement metric of the variation of function derived from the analytical performance boundary approximation [12]. A schematic depiction of our active learning technique is shown in Figure 1.

Our suggesting methods are basically derived from the augmentation-based consistency estimation. For the training step of active learning, we devised augmentation-based consistency measurement (estimation) for regularization loss. We will call it as consistency-based regularization or augmentation-based regularization for convenience. For the candidate sampling step of active learning, we suggest augmentation-based uncertainty measurement, briefly saying augmentation-based uncertainty or uncertainty measure.

The technical contributions of the active learning method using the consistency-based regularization are as follows.

- We apply the augmentation-based measurement metric for the candidate sampling step of active learning scenarios as an uncertainty measure. To naturally apply augmentation-based techniques to active learning, we propose a method of measuring the variation of function derived from the analytical learning theory. By doing so, our algorithm selects the data which is hard for the current model to ensure consistency.

- Consistency-based regularization can be adapted to the training step and contributed to active learning as a regularizer. In particular, the regularization loss combined with any uncertainty measurement (existing and our suggesting) techniques yielded an improved generalization performance.

- Through deep active learning with our suggesting consistency-based methods, we show the steady improvement of generalization performance that is not significantly affected by dataset and budget scenario changes in the image classification.

2. Background

Active Learning. Active learning in machine learning has been proposed mainly as a method of selecting the samples to be labeled by measuring the uncertainty of the unlabeled samples. Representative uncertainty measurement methods use the entropy for the probability distribution of the classifier as a processed input feature [9] or the margin between the input feature and the hypothesis function [1], or the margin between the feature vectors [15]. These uncertainty measurement methods are also used for active learning using CNNs, resulting in improved performance in image recognition problems. [11] used Gaussian process models for expected model output changes for active learn-
ing, and [5] applied Bayesian inference based on dropout to use probabilistic approaches using Bayesian prior to active learning. Efforts have also been made to improve the efficiency of active learning by using ensemble approaches [2] to apply a Query-by-Committee strategy to deep neural networks or by predicting expected losses with submodules that rely on recognition modules for target tasks [20].

However, many active learning techniques in the image recognition field using deep neural networks show performance sensitivity to the dataset and budget scenarios, and it has been reported that the actual performance is difficult to reproduce [17]. Active learning, using our proposed consistency-based regularization, is more deeply considered in the learning methodology for deep neural networks than the previous active learning techniques, and provides consistent performance improvements that are independent of the dataset and budget scenarios.

**Data Augmentation-based Regularization.** Data augmentation-based regularization techniques help to improve generalization performance in deep neural networks training. Unlike the explicit regularization methods such as weight decay which is applied with a statistical learning approach, data augmentation-based regularization is aimed at preventing overfitting through transformation on input data during training [22]. Data augmentation-based regularization techniques are applied mainly using transformation methods but recently proposed data augmentation-based regularization techniques [4, 21, 22] have used intentional data mixing and soft labeling [8] to improve the generalization performance. Active learning, which requires efficient learning using such a limited amount of training data, is inextricably linked with data augmentation-based regularization. We reformatted both the cutout [4] and cutmix [21] among the data augmentation-based regularization techniques to assist in the active learning cycle. Using redefined data augmentation-based regularization, we could apply both the process of selecting the data to be labeled and the process of training the classifier in each active learning cycle.

**Analytic Learning Theory.** Analytical learning theory is a measure-theoretic learning approach for machine learning suggested in [12]. It is based on a non-statistical method that minimizes model assumptions about the data and uses the characteristics of the data according to empirical observations. Because model training using deep neural networks often does not follow known probability distributions or model assumptions, various analyses based on empirical observations have been reported [23]. Our proposed active learning method is based on [12] that transforms a database-based regularization method into a consistency-based loss by analyzing the generalized error bound of deep neural networks through the analysis learning theory. We extend the measurement of the variation of function derived from the analytical learning theory from dual-cutout [12] to k-cutout and apply it to each active learning cycle. To adopt cutmix augmentation for active learning, we add some constraints on the definition of cutmix strategy to preserve the data semantics. For the training cycle of deep active learning, we redefine the cutmix method using soft labels into the consistency-based regularization loss according to the analytical learning theory. In addition, we convert cutmix regularization into the entropy-based uncertainty measurement technique for the candidate sampling step of active learning. By redefining the data augmentation-based techniques, we can apply them to active learning scenarios and develop training strategies that are less affected by dataset and budget constraints.

3. Deep Active Learning with Consistency-based Regularization

This section describes active learning using our proposed consistency-based techniques. First, we define problems and notations for active learning scenarios and then explain how the estimation of variation of function in active learning scenarios is related to the generalization of deep neural networks. We also describe an approach to apply cutout and cutmix as uncertainty measures, using representative data augmentation, to active learning scenarios. In addition, we provide a semantic visualization of how consistency-based loss with data augmentation affects the total loss function during training, thus explaining why the proposed training strategy shows improved performance.

3.1. Background and problem formulation

To define an active learning scenario in a t cycle for sample set $S = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ for data and label pairs, it consists of labeled data $L_{b,t}$ and unlabeled data $U_{N-b,t}$ in the entire dataset $D_N = L_{b,t} + U_{N-b,t}$. At this time, the initial training samples are fixed to b uniform samples $L_{b,0}$. Active learning consists of $b(t+1)$ labeled data and $N-b(t+1)$ unlabeled data in $D_N$ according to cycle $t$ and budget $b$. In this case, the active learning process is performed by selecting $b$ data to be labeled in the next cycle among the unlabeled data $U_{N-b,t}$ in every training cycle. Thus, when selecting $b$ data to be labeled in a particular cycle $t$ of active learning, the following conditions must be met for the classifier $h(x, y; \theta)$ to be trained:

$$\arg\max_{L_{b,t+1} - L_{b,t}, \theta_{t+1}} E[h_{L_{b,t+1}}(x, y; \theta_{t+1})] - E[h_{L_{b,t}}(x, y; \theta_{t})],$$

(1)

where $E$ is the generalization error of classifier $h$ and $h_{L_{b,t}}(x, y; \theta_{t})$ is a classifier with parameter $\theta_{t}$ trained in $t$ cycles using labeled data $L_{b,t}$. According to Equa-
tion (1), active learning is a problem of finding the subset \( \mathcal{L}_{b,t+1} - \mathcal{L}_{b,t} \subseteq \mathcal{U}_{N-b,t} \) of the data having the largest generalization gap between the previous cycle and the current cycle and the classifier parameter \( \theta_{t+1} \). To satisfy this condition, we must define a selection strategy for finding good \( \mathcal{L}_{b,t+1} - \mathcal{L}_{b,t} \) and a learning strategy for finding good \( \theta_{t+1} \) at cycle \( t+1 \).

### 3.2. Variation of function estimation via analytical learning theory

We used an approach that utilizes the analytical learning theory for the generalized error bound in machine learning proposed in [12], to simultaneously achieve two active learning goals given in Equation (1). The generalized error bound based on the analytical learning theory demonstrated in [12] is given by:

\[
E[h_{\mathcal{L}}(x, y; \theta)] - \hat{E}[h_{\hat{\mathcal{L}}}(x, y; \hat{\theta})] 
\leq \sum_{y \in \mathcal{Y}} c_2 p(y) \sigma(f_y) \sqrt{\frac{d_z}{|\hat{\mathcal{L}}_z|}} + \hat{E}[h_{\hat{\mathcal{L}}}(x, y; \hat{\theta})] \sqrt{\frac{\log(2/\delta)}{2|\mathcal{L}_{x|y}|}},
\]

(2)

where \( \mathcal{L} \) is a theoretical pair of labeled data that can minimize generalization errors and \( \hat{\mathcal{L}} \) is a given set of labeled data for actual training. \( E[h_{\mathcal{L}}(x, y; \theta)] - \hat{E}[h_{\hat{\mathcal{L}}}(x, y; \hat{\theta})] \) is a generalization gap for \( \mathcal{L} \) and the labeled subset \( |\hat{\mathcal{L}}_z| \) is given by \( |\hat{\mathcal{L}}_z|/|\hat{\mathcal{L}}_x| \). In this case, \( d_z \) given in the dimension of the \( z \)-level hidden layer and the constant \( c_2 \) are developed in Proposition 2 of [12]. The adjustable term for the generalization error boundary during training is the amount of \( \sigma(f_y) \) variation of function. According to Equation (2), minimizing \( \sigma(f_y) \) minimize the upper boundary of generalization error. [12] proposed a consistency-based regularization loss using dual-cutout augmentation on input data [4] to reflect \( \sigma(f_y) \) during training and is defined as:

\[
L_{\text{reg}}(x, \theta) = \int_{(x_1^{co}, x_2^{co})} \| h(x_1^{co}; \theta) - h(x_2^{co}; \theta) \|_2^2 dP(x_1^{co}, x_2^{co}|x),
\]

(3)

where \( P(x_1^{co}, x_2^{co}|x) \) is defined as two random cutouts for one input data. We have modified and redefined the cutout and cutmix to fit the active learning scenario using the definition of consistency-based regularization loss in Equation (3). At the same time, the regularization technique using data augmentation was scored and applied to the active learning cycle with uncertainty for the selection of data to be learned in the next cycle.

### 3.3. Deep active learning with \( k \)-cutout

First, we extend the random cutout generation from dual to \( k \) in order to use cutout for the uncertainty measurement method of dual cutout proposed in [12]. The \( k \)-cutout for uncertainty measurement is then defined as:

\[
U(x, \theta) = \int_{(x_1^{co}, x_2^{co})} \| h(x_1^{co}; \theta) - h(x_2^{co}; \theta) \|_2^2 dP(x_1^{co}, x_2^{co}|x)
\]

\approx \frac{1}{|K_{co}|} \sum_{i=1}^{K_{co}} (h(x_i^{co}; \theta_i) - h(x_{j}^{co}; \theta_i))^2, (i \neq j),
\]

(4)

where \( |K_{co}| \) is the number of pairwise cases when \( K_{co} \) random cutout images are generated and \( x_i^{co} \) is input data applying an arbitrary cutout to the \( i^{th} \) sample. \( K_{co} \) random cutout images were generated for \( k \)-cutout and the mean value of the pairwise distance for each inference result was defined as uncertainty. Based on the estimated uncertainty, we trained the data with the large amount of variation of function first and expected the effect of minimizing the generalization error boundary to a greater extent than other unlabeled data. At the same time, it was expected to show good generalization performance for training after data selection in combination with consistency-based regularization loss. Consistency-based regularization loss using cutout when training about the current cycle is defined as follows.

\[
L_{\text{co}}(x, \theta) = \frac{1}{M} \frac{1}{|K_{co}|} \sum_{i=1}^{K_{co}} (h(x_i^{co}; \theta_{t+1}) - h(x_{j}^{co}; \theta_{t+1}))^2
\]

\[+ \frac{1}{|K_{co}|} \sum_{i=1}^{K_{co}} CE(h(x_i^{co}; \theta_{t+1}), y_i), (i \neq j),
\]

(5)

where \( M \) represents the size of the mini-batch and \( CE \) represents the cross-entropy function. In Equation (5), two regularization terms are applied to efficiently reflect the variation of function through cutout during training. The first term is a regularization term for minimizing mean squared error (MSE) for different \( k \)-cutout samples. The network must be trained in such a direction as to minimize the variation of function for \( K_{co} \) number of cutout samples. The second term is the cross-entropy output with the ground truth label for the \( k \)-cutout samples, which includes a condition to minimize the amount of variation of function and deduce the correct answer with the corrupted image. The total loss using \( k \)-cutout is given by \( L_{\text{total}} = L_{\text{ce}} + L_{\text{co}} \), where \( L_{\text{ce}} \) is cross-entropy loss for the target task.

### 3.4. Deep active learning with cutmix

In addition to the cutout, we applied the cutmix technique [21] to measure uncertainty and loss of consistency-based regularization for active learning scenarios, since cutmix is known to improve the generalization performance by data augmentation-based regularizations and soft label. To
measure the uncertainty based on cutmix, we need $K_{cm}$ different cutmix processes on one unlabeled sample and calculate the entropy of inference output $h(x_i; \theta_t)$. At this time, we choose an unlabeled sample that shows unstable inference results for several cutmix images with different semantics, expecting that it can stabilize the variation of the function of the $h(x_i; \theta_t)$ by adding it into the labeled set.

$$U(x, \theta) = -\frac{1}{K_{cm}} \sum_{i=1}^{K_{cm}} h(x_i; \theta_t) \log h(x_i; \theta_t). \quad (6)$$

Meanwhile, the consistency-based regularization loss with cutmix is defined as follows:

$$L_{cm}(x, \theta) =$$

$$\frac{1}{M} \left( \frac{1}{K_{cm}C_2} \sum_{i=1}^{K_{cm}C_2} (h(x_i^{cm}; \theta_{t+1}) - h(x_j^{cm}; \theta_{t+1}))^2 \right) +$$

$$\frac{1}{K_{cm}} \sum_{i=1}^{K_{cm}} CE(h(x_i^{cm}; \theta_{t+1}), y_i), (i \neq j), \quad (7)$$

where $K_{cm}C_2$ is the number of pairwise cases for $K_{cm}$ arbitrary cutmix images of the $i^{th}$ sample. The consistency-based regularization loss using cutmix is defined as the linear combination of the distance metric and the cross-entropy output, in the same way as the cutout-based regularization loss. The reason for not including the entropy term used to measure uncertainty in the cutmix-based regularization loss, is to avoid using additional balance parameters or normalization constants for the total loss. In this case, the total loss is defined as $L_{total} = L_{ce} + L_{cm}$. Finally, the total loss of both cutout and cutmix is given by $L_{total} = L_{ce} + L_{co} + L_{cm}$. The gradient of the backpropagation for the total loss during training in $t + 1$ cycle is then obtained by $\nabla L_{total} = \nabla L_{ce} + \nabla L_{co} + \nabla L_{cm}$.

The semantic visualization of the backpropagation process in Figure 2 shows the role of our consistency-based regularization losses. It depends on the derivative of the total loss. In Figure 2(a), the length of each arrow indicates the magnitude of the gradient. It means how much each loss could affect updating model parameters. In the case of cutmix, the cross-entropy loss $L_{cm,ce} = \frac{1}{K_{cm}} \sum_{i=1}^{K_{cm}} CE(h(x_i^{cm}; \theta_{t+1})$ is derived from differences comparing with correct labels. It has a larger magnitude of gradient than MSE loss $L_{cm,mse} = \frac{1}{K_{cm}C_2} \sum_{i=1}^{K_{cm}C_2} (h(x_i^{cm}; \theta_{t+1}) - h(x_j^{cm}; \theta_{t+1}))^2$, and it is occurred from differences between the inference output of augmented data from the same sample. The red arrow shows an example of the gradient magnitude that occurred in the sample with large uncertainty, and the black one indicates another sample with small uncertainty.

In Figure 2(b), there are examples of the directional derivative of each loss for low uncertainty samples. The samples in left-hand side of the dotted line indicates low uncertainty cases. Since samples with low uncertainty has a similar softmax output to the input sample even after data corruption, it’s derivative has a similar direction to the target loss $\nabla L_{ce}$ proceeds. In this situation, the direction of $\nabla L_{ce} + \nabla L_{cm}$ is much like the input sample’s own $\nabla L_{ce}$. As a result, the gradient of total loss $\nabla L_{total}$ has a similar effect to taking a larger step in a direction similar to the direction of inclination of $\nabla L_{ce}$ to proceed. In other words, regularization loss is less likely to diversify the information that the model attains. This is likely to help with fast convergence similar to training with a large learning rate, but it is difficult to help find a solution with good generalization performance. Conversely, for input samples with large uncertainties (examples to the right of the dashed line), the derivative of each loss is more likely to be given a different direction to the total loss. As a result, $\nabla L_{total}$ is likely to have a gradient fall in a direction different from the direction in which $\nabla L_{ce}$ proceeds, which means that there is a high possibility of performing a parameter update that is advantageous to finding a point potentially having high generalization performance. For this reason, it is presumed that training methodologies, including data augmentation-based uncertainty estimations and consistency-based regul-
Table 1. Active learning results using the FashionMNIST dataset \((b = 300)\). Each cell represents the performance of a cycle of \(t\) for each active learning technique, which is measured 3 times. Values shown in red indicate improved accuracy compared to random samples, while values in bold blue indicate the highest accuracy of any active learning strategy. The last row shows the difference in accuracy between random samples and the active learning strategy with the greatest performance gains.

| uncertainty/loss | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------------|---|---|---|---|---|---|---|---|---|----|
| random/task      | 65.51 | 71.84 ± 0.99 | 80.97 ± 0.44 | 84.33 ± 0.47 | 85.34 ± 0.98 | 86.12 ± 0.19 | 86.99 ± 0.34 | 86.66 ± 0.22 | 87.61 ± 0.66 | 86.65 ± 0.48 |
| cutout/task      | 63.51 | 74.29 ± 0.49 | 79.01 ± 0.44 | 81.36 ± 0.17 | 85.1 ± 0.1 | 86.09 ± 0.25 | 89.94 ± 0.13 | 87.53 ± 0.13 | 87.06 ± 0.29 | 87.62 ± 0.06 |
| cutmix/task      | 63.51 | 73.61 ± 2.5 | 81.82 ± 0.56 | 82.07 ± 1.44 | 83.35 ± 0.53 | 86.12 ± 0.08 | 86.63 ± 0.59 | 87.17 ± 0.67 | 87.53 ± 0.54 | 88.35 ± 0.41 |
| entropy/task     | 63.51 | 76.86 ± 0.26 | 81.96 ± 0.27 | 85.36 ± 0.47 | 86.34 ± 0.48 | 87.74 ± 0.66 | 89.66 ± 0.4 | 90.7 ± 0.2 | 91.03 ± 0.16 | 91.91 ± 0.01 |
| margin/task      | 63.51 | 77.53 ± 0.62 | 83.06 ± 0.66 | 86.35 ± 0.37 | 87.95 ± 0.29 | 89.11 ± 0.09 | 90.1 ± 0.34 | 90.42 ± 0.14 | 90.49 ± 0.17 | 91.29 ± 0.25 |
| cutout/task+eco  | 63.51 | 79.71 ± 0.91 | 84.81 ± 0.41 | 86.75 ± 0.56 | 87.46 ± 0.73 | 89.9 ± 0.36 | 89.21 ± 0.52 | 89.21 ± 0.03 | 90.16 ± 0.07 | 90.09 ± 0.11 |
| cutmix/task+eco  | 63.51 | 81.5 ± 1.17 | 85.25 ± 0.71 | 87.84 ± 0.5 | 88.38 ± 0.21 | 89.32 ± 0.5 | 90.0 ± 0.14 | 90.09 ± 0.36 | 90.64 ± 0.56 | 90.85 ± 0.12 |
| entropy/task+eco | 63.51 | 80.06 ± 0.14 | 84.66 ± 0.23 | 86.42 ± 0.25 | 87.48 ± 0.38 | 88.34 ± 0.32 | 89.0 ± 0.15 | 89.1 ± 0.08 | 89.64 ± 0.38 | 90.84 ± 0.43 |
| cutmix/task+eco  | 63.51 | 80.19 ± 0.85 | 85.98 ± 0.69 | 87.02 ± 0.56 | 88.18 ± 0.54 | 89.50 ± 0.41 | 89.28 ± 0.21 | 90.44 ± 0.29 | 90.59 ± 0.31 | 91.12 ± 0.13 |
| cutout/task+eco+cm | 63.51 | 80.4 ± 0.74 | 85.0 ± 0.36 | 86.91 ± 0.44 | 87.87 ± 0.21 | 89.44 ± 0.37 | 89.55 ± 0.26 | 89.42 ± 0.25 | 90.11 ± 0.11 | 90.45 ± 0.11 |
| cutmix/task+eco+cm | 63.51 | 81.02 ± 0.73 | 85.96 ± 0.52 | 87.73 ± 0.79 | 89.6 ± 0.02 | 90.08 ± 0.41 | 90.54 ± 0.5 | 91.19 ± 0.3 | 91.55 ± 0.46 | 91.47 ± 0.12 |
| maximum gap      | 0 | 3.66 | 5.91 | 7.91 | 4.35 | 3.96 | 3.55 | 4.53 | 4.37 | 3.97 |

Table 2. Active learning results using consistency-based regularization loss in the FashionMNIST dataset \((b = 300)\). Combined with various uncertainty measures, the results show improved performance.

| uncertainty/loss | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------------|---|---|---|---|---|---|---|---|---|----|
| random/task      | 65.51 | 71.84 ± 0.99 | 80.97 ± 0.44 | 84.33 ± 0.47 | 85.34 ± 0.98 | 86.12 ± 0.19 | 86.99 ± 0.34 | 86.66 ± 0.22 | 87.61 ± 0.66 | 86.65 ± 0.48 |
| entropy/task     | 63.51 | 81.89 ± 0.61 | 87.53 ± 0.52 | 88.69 ± 0.55 | 89.09 ± 0.04 | 91.44 ± 0.19 | 91.64 ± 0.15 | 92.09 ± 0.13 | 92.24 ± 0.37 | 92.56 ± 0.23 |
| margin/task      | 63.51 | 81.29 ± 0.67 | 84.94 ± 0.41 | 85.92 ± 0.98 | 87.44 ± 0.11 | 89.63 ± 0.68 | 89.96 ± 0.44 | 90.22 ± 0.48 | 89.85 ± 0.44 | 91.04 ± 0.35 |
| entropy/task+eco | 63.51 | 78.98 ± 1.46 | 84.94 ± 0.66 | 85.96 ± 0.17 | 87.68 ± 0.41 | 89.78 ± 0.46 | 90.18 ± 0.2 | 90.80 ± 0.29 | 91.27 ± 0.41 | 91.85 ± 0.15 |
| maximum gap      | 0 | 5.89 | 7.79 | 5.43 | 5.54 | 5.32 | 4.73 | 5.43 | 4.37 | 3.97 |

Figure 3. Active learning performance on FashionMNIST with different strategies and regularization losses. For all graphs, the red line shows the performance of random sampling without any regularizer, (a) shows the performance of active learning using uncertainty sampling based on cutout, cutmix, entropy and margin in clockwise direction, and (b) shows the active learning performance using no loss, \(L_{co}, L_{cm}\), and \(L_{co} + L_{cm}\) based uncertainty sampling in clockwise direction, respectively.

4. Experiment Results

In order to verify the performance of active learning using augmentation-based methods, we conducted experiments with CNNs in various budget scenarios with multiple image classification datasets. For this purpose, we divided the experiments into three cases to analyze the effect of our proposed methods on active learning. The first is whether data augmentation-based regularization can improve performance in active learning scenarios when uncertainty and regularization loss are utilized together. For this, we tested whether the consistency-based regularization loss with arbitrary uncertainty measures can lead to performance improvement. The second is a comparative analysis of the proposed active learning technique for different datasets. By using representative datasets in an image classification problem, we confirmed that our methods can achieve consistent performance improvement in different datasets. Finally, we conducted active learning experiments on multiple budget scenarios.\(^{17}\) showed that the results of active learning experiments using deep neural networks can be very sensitive to budget scenarios. We verified the robustness of our active learning methods on multiple budgets in the controlled experimental design. For the overall performance report, please refer to Figure 4.

Dataset. We conducted a series of experiments to validate multiple active learning scenarios of representative image recognition tasks. FashionMNIST\(^{19}\), CIFAR10\(^{13}\), CIFAR100\(^{13}\), and STL10\(^{3}\) were used for the experiment. FashionMNIST consists of black and white images of 10 classes of clothing, each of which has a size of 28 by 28. The dataset consists of 60,000 training images and 10,000 test images. The CIFAR10 consists of 32 by 32 color images of 10 natural scenes, divided into 50,000 training images and 10,000 test images. CIFAR100 is designed to solve fine-grained image classification problems by subdividing data into 100 classes with detailed class information about CIFAR10. The STL10 is a dataset designed to solve the problem of classifying large amounts.
of test data with a relatively small number of labeled images, with ten classes for color images of size 96 by 96. Each class consists of 500 training samples and 800 test samples.

**Training Detail.** As a data transformation to prevent overfitting during the training process, the random flip was performed after applying the 4-pixel padding that is common to all datasets. We use ResNet18 [6] as a training model for image classification. The initial labeled data $L_{b,0}$ for active learning is fixed under the same configuration obtained with uniform samples, and initial model $h_{L_{b,0}}(x; \theta_0)$ is set to the same weight initialization. In each $t + 1$ active learning cycle, $b$ number of samples were selected after uncertainty measurements on the unlabeled sample set $U_{N-b,t}$ to find label candidate samples using the current model $h_{L_{b,t}}(x; \theta_t)$. After selecting the candidate samples, samples were added to $L_{b,t+1}$ to perform training in $t + 1$ cycles starting from $L_{b,t+1}$ label samples with the same weight initial value as the model $h_{L_{b,t}}(x; \theta_t)$. In all training, we use a stochastic gradient descent (SGD) optimizer, with an initial learning rate 0.1, a momentum 0.9, a weight decay $5e^{-4}$, and a mini batch size 128, and total 200 epochs for each active learning cycle. The learning scheduler applied a scale of 0.1 at 160 epochs for all datasets except CIFAR100, and 0.2 scales at 60, 120, and 160 epochs for training using the CIFAR100 dataset.

4.1. Deep active learning using data augmentation-based regularization

To verify our suggesting active learning methods, we measured the performance according to the uncertainty measurement method. To check whether our methods can be extended to existing active learning uncertainty methods, we used entropy- and margin-based uncertainty measures with the same CNN model as the baseline learning method. Although a variety of uncertainty measures have been proposed [15] in addition to the two approaches, recent studies have shown sensitive results as changes in the dataset and budget scenarios occur [17]. Furthermore, in the case of [20] that takes advantage of a submodule attached to the deep neural networks, it is difficult to accurately compare the performance because of additional parameters of the submodule.

The results of active learning using cutout and cutmix under a specific budget scenario and dataset are shown in Table 1 and Table 2. Since the proposed approach is divided into the data augmentation technique in the uncertainty measurement and regularization loss in the training process in the active learning scenario, the performance is divided into the sampling and learning strategy. In Table 1, we can verify that the proposed active learning technique shows better performance than the entropy- and margin-based active learning techniques. The performance varia-
tion of the proposed consistency-based regularization loss combined with other uncertainty measures is shown in Table 2. In addition to the data augmentation-based uncertainty measurement method proposed here, it is evident that the method shows good performance in combination with arbitrary uncertainty measurement methods.

In Figure 3, graphical examples of performance variation with a particular budget scenario \( (b = 300) \) in the FashionMNIST dataset are shown. In all graphs, the solid black line represents the result of training with target loss only on random sampling. The graphs in Figure 3 compare with several active learning scenarios with fixed uncertainty measurements. In the case of using the uncertainty technique without the consistency-based regularization loss, the performance of each active learning method is improved not that much. However, when the regularization loss was applied together during the training, it was confirmed that the performance steadily improved in all the uncertainty measurement-based active learning. In particular, the results of the bottom row show that the performance improvement is also shown for the uncertainty measurement method using entropy and margin, which means that other existing active learning methods can be combined with the regularization losses.

4.2. Robustness for datasets

To verify the effectiveness of the proposed active learning method in various datasets, image classification experiments were conducted on FashionMNIST, CIFAR10, CIFAR100, and STL10. In Figure 4 Exp A (first row), the average performances of the entire active learning cycle are visualized to show the performance gain efficiently. We divide the active learning scenarios into 4 approaches: ‘baseline’ which is random sampling with no consistency-based regularization loss, ‘only strategies’ which is our proposed uncertainty method or existing rule with no regularization loss, ‘only losses’ which is random sampling with our regularization losses, and ‘both’ which is the combination of uncertainty methods with our regularization losses. Each colored-circle indicates the average accuracy in each setting. Our experiments are conducted among the following combinations: random sampling, cutout, cutmix, entropy sampling, and margin sampling as possible strategies, and no regularization, cutout, cutmix, and cutout with cutmix regularization as possible losses.

First, Figure 4 Exp A visualize the average performance of each active learning settings in all dataset. It shows that consistency-based regularization loss combined with uncertainty measures consistently enhances the performance in all datasets. Secondly, only the regularization loss or uncertainty measure alone may not effective due to the characteristics of each dataset, together exploits the natural synergy between their contribution that is explained in our schematic depiction in Figure 1. It is especially noteworthy that CIFAR100 and STL10, which have a relatively small number of images per class, do not show any significant performance improvement when only regularization loss is combined with random sampling, but show high performance in combination with uncertainty measurement.

4.3. Robustness for budgets

We conducted experiments based on the active learning scenarios with different budgets in a specific dataset to verify various budgets for the proposed active learning technique. For this purpose, experiments were conducted in \( b \in \{500, 1000, 2000\} \) in the CIFAR10 dataset. The results of active learning under different budgets are shown in Figure 4 Exp B. Starting from a small amount of labeled data to a sufficient scale of selected data, we can see the performance gain on average. In other words, if a certain amount of label data is secured, there are overall performance increases in all cases especially when arbitrary data sampling budget rule is combined with our suggested data augmentation-based regularization loss.

5. Conclusions

We proposed active learning methodologies using the augmentation-based consistency estimation derived from the analytical learning theory. Based on the analytical learning theory, it can be confirmed that the variation of function obtained from the observed data can have a substantial influence on the generalization error of the learning model. By adopting our augmentation-based consistency estimation methods on each active learning cycle, we can achieve consistent performance improvement and achieve a high-performance improvement in combination with the previously proposed uncertainty measurement methods. At the same time, we redefined the active learning scenario related to deep learning and visually explained how the proposed methods work on our definition.

Nevertheless, there are limitations in experimenting only with the image classification problem. The proposed active learning methodology needs to be applied to various image recognition tasks, and additional experiments are needed to see if the same conclusion can be reached for large data. We believe that the analysis of data using analytical learning theory can be a good starting point for accessing various learning methodologies. In particular, it is expected that it can be effectively used for semi- or self-supervised learning that uses unlabeled data, and it can be useful for various learning methodologies that utilize unlabeled data.

References

[1] M. F. Balcan, A. Broder, and T. Zhang. Learning theory. *Margin based active learning*, pages 35–50, 2007.
[2] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler. The power of ensembles for active learning in image classification. In Proc. of CVPR, 2018.
[3] A. Coates, H. Lee, and A. Y. Ng. An analysis of single layer networks in unsupervised feature learning. In Proc. of AISTATS, 2011.
[4] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. In arXiv:1708.04552, 2017.
[5] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In Proc. of ICML, 2017.
[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proc. of CVPR, 2016.
[7] A. Hernandez-Garcia and P. Konig. Data augmentation instead of explicit regularization. In arXiv:1806.03852, 2019.
[8] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In Proc. of NIPS, 2014.
[9] A. Holub, P. Perona, and M.C. Burl. Entropy-based active learning for object recognition. In Proc. of CVPRW, 2008.
[10] P. Hu, Z. C. Lipton, A. Anandkumar, and D. Ramanan. Active learning with partial feedback. In Proc. of ICLR, 2019.
[11] C. Kading, E. Rodner, A. Freytag, O. Mothes, B. Barz, and J. Denzler. Active learning for regression tasks with expected model output changes. In Proc. of BMVC, 2018.
[12] K. Kawaguchi, Y. Bengio, V. Verma, and L. P. Kaelbling. Generalization in machine learning via analytical learning theory. In arXiv: 1802.07426, 2018.
[13] A. Krizhevsky. Learning multiple layers of features from tiny images. In Tech. Report, 2009.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Proc. of NIPS, 2012.
[15] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In Proc. of ICLR, 2018.
[16] Burr Settles. Active learning literature survey. In Technical Report, 2010.
[17] O. Simeoni, M. Budnik, Y. Avrithis, and G. Gravier. Re-thinking deep active learning: Using unlabeled data at model training. In arXiv:1911.08177, 2019.
[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. of ICLR, 2015.
[19] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In arXiv:1708.07747, 2017.
[20] D. Yoo and I. S. Kweon. Learning loss for active learning. In Proc. of CVPR, 2019.
[21] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proc. of ICCV, 2019.
[22] C. Zhang, S. Bengio, M. Hardt, and O. Vinyals B. Recht and. Understanding deep learning requires rethinking generalization. In Proc. of ICLR, 2017.
[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In Proc. of ICLR, 2018.