Zero-Shot Knowledge Distillation Using Label-Free Adversarial Perturbation With Taylor Approximation

KANG IL LEE, SEUNGHYUN LEE, (Associate Member, IEEE),
AND BYUNG CHEOL SONG, (Senior Member, IEEE)
Department of Electronic Engineering, Inha University, Incheon 22212, South Korea
Corresponding author: Byung Cheol Song (bcsong@inha.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant by the Korean Government through the Ministry of Science and Information & Communications Technology (MSIT), in part by the Artificial Intelligence Convergence Research Center, Inha University under Grant 2020-0-01389, and in part by the Engineering Development Research Center (EDRC) through the Ministry of Trade, Industry, and Energy (MOTIE) under Grant P0008475.

ABSTRACT
Knowledge distillation (KD) is one of the most effective neural network light-weighting techniques when training data is available. However, KD is seldom applicable to an environment where it is difficult or impossible to access training data. To solve this problem, a complete zero-shot KD (C-ZSKD) based on adversarial learning has been recently proposed, but the so-called biased sample generation problem limits the performance of C-ZSKD. To overcome this limitation, this paper proposes a novel C-ZSKD algorithm that utilizes a label-free adversarial perturbation. The proposed adversarial perturbation derives a constraint of the squared norm of gradient style by using the convolution of probability distributions and the 2nd order Taylor series approximation. The constraint serves to increase the variance of the adversarial sample distribution, which makes the student model learn the decision boundary of the teacher model more accurately without labeled data. Through analysis of the distribution of adversarial samples on the embedded space, this paper also provides an insight into the characteristics of adversarial samples that are effective for adversarial learning-based C-ZSKD.

INDEX TERMS
Zero-shot learning, knowledge distillation, adversarial learning.

I. INTRODUCTION
With the advent of effective solutions [1], [2] to the gradient vanishing problem, deep neural networks that provide high recognition performance have been developed rapidly. However, the still high computational cost of deep neural networks hinders the practical use of deep learning. Hinton et al. firstly introduced the concept of knowledge distillation (KD) to effectively lighten neural networks [3]. KD is a technique that transfers knowledge from large networks that perform similar tasks to relatively small networks. So KD not only allows small networks to overcome the limitations of training, but ultimately tries to have the same performance as large networks. Conventional KD techniques implicitly assume that training data is always available. However, it is often impossible to access training data due to legal factors such as personal information protection. In this case, conventional KD techniques cannot be used. To overcome this limitation, zero-shot KD (ZSKD) algorithms have been recently proposed. One of the ZSKD approaches is the generalized ZSKD (G-ZSKD) [4], [5] that utilizes public data or meta data instead of original training data. The other one is complete ZSKD (C-ZSKD) [6], [7] that does not use external training data at all. A G-ZSKD method assumes that information on layer-wise activation when training the teacher model is meta data [4]. The meta data is converted into knowledge which is used for training the student model. C-ZSKD is a method of creating the training data instead of training external datasets such as meta data and public datasets. Nayak et al. generated a pseudo dataset similar to the dataset used to learn the teacher model using a parametric distribution, and then trained the student model [6]. Micaelli et al. adopted the adversarial learning (AL) of the generator and student model to transfer the decision boundary information of the teacher model to
the student model [7]. Micaelli’s method is relatively easy to apply to various computer vision tasks such as segmentation and object detection because it has no statistical restriction, unlike Nayak’s.

However, C-ZSKD based on AL suffers from the so-called biased sample generation problem in which adversarial samples produced by the generator are mapped only to particular decision boundaries of the teacher model. Figure 1 explains the phenomenon that the output distribution of the teacher model receiving adversarial samples is biased to a specific class. This prevents the student model from accurately learning the teacher model’s decision boundary through adversarial samples. To solve this problem, conventional adversarial attack techniques based on adversarial samples adopted a method of giving perturbation in label space [8-10]. However, the conventional techniques require labeled data and cannot be used in zero-shot learning. To solve this problem, we propose label-free adversarial perturbation. Specifically, by estimating the perturbed sample distribution using Taylor approximation, sample variance is improved without labels. In addition, to guarantee convergence of learning perturbed adversarial samples, a constraint to have a tight upper bound is presented. As a result, effective learning is realized because adversarial samples generated by the proposed method more clearly provide the decision boundary of the teacher network.

The contribution points of this paper are summarized as follows.

- To solve an inherent biased sample generation problem of AL-based C-ZSKD, we propose a method to increase the variance of the adversarial sample distribution by using the convolution of probability distributions and Taylor series approximation.
- By analyzing the distribution of adversarial samples in the embedding space, this paper provides an insight into the characteristics of adversarial samples that are useful for AL-based C-ZSKD.

II. BACKGROUND

A. COMPLETE ZERO-SHOT KNOWLEDGE DISTILLATION

Unlike G-ZSKD, C-ZSKD can be said to be a highly scalable method because it operates even in an environment where training data is completely blocked. C-ZSKD is categorized into two approaches. The first approach is focusing on producing information of training data from the teacher model (T) by generating a pseudo dataset similar to the training data of T. For example, Nayak et al. estimated the number of classes from the weight of the final layer of T [6]. Then, assuming that the label of each class follows a Dirichlet distribution (D), the concentration parameter of D was derived from the weight W. Next, the pseudo labels (ŷ) were sampled from D, and the corresponding pseudo image (x*) was generated according to Eq. (1).

\[ x^* = \arg\min_x L_{CE}(T(\tilde{x}), \hat{y}) \quad \text{(1)} \]

\[ L_{CE}(\cdot, \cdot) \] in Eq. (1) indicates cross entropy loss. Then, the pseudo dataset is generated by repeating the above-mentioned process. Finally, the student model (S) was trained using the pseudo dataset and the conventional KD technique [3]. The second approach is to create adversarial samples to transfer decision boundary information of T to S [7]. Specifically, AL was applied to the generator (G) and S according to Eqs. (2) and (3).

\[ W_G = \arg\max_{W_G} D_{KL}(T(G(z)) \mid \mid S(G(z))) \quad \text{(2)} \]

\[ W_S = \arg\min_{W_S} D_{KL}(T(G(z)) \mid \mid S(G(z))) \quad \text{(3)} \]

Here, \( W_G \) and \( W_S \) indicate the weights of G and S, respectively. Also, \( D_{KL} \) means Kullback-Leibler (KL) divergence, and z is the latent code sampled from \( N(0, I) \). The AL was performed as follows. First, according to Eq. (2), G is trained so that S and T output different logits. Next, S receives the decision boundary information from T using the adversarial samples as in Eq. (3). Here, since the pseudo dataset generation method [6] must choose a parametric probability distribution suitable for each task, S is subject to statistical restrictions by parameters. So this method is very hard to fully reflect the actual data distribution. On the other hand, the AL-based method [7] need not select the parametric probability distribution for each task. Above all, this method has an advantage to reflect the actual data distribution because it can learn the probability distribution of the adversarial samples located near the decision boundary of the teacher model without statistical limitations. Therefore, this paper adopts the AL-based C-ZSKD approach for further effective training of the student model.

B. ADVERSARIAL PERTURBATION

Since adversarial perturbation (AP) had been introduced in [9] as a means to attack neural networks, a lot of AP generation techniques have been developed. In general, if arbitrary data is input to a neural network together with AP, the data can move from its original classification region to another classification region [10]. Based on this property, Heo et al. generated adversarial samples using AP, and then applied the produced adversarial samples to KD [11]. Since the adversarial samples generated by [11] exist near the decision boundary of the teacher model, the decision boundary information of the teacher model could be easily transferred to the student model. As a result, Heo et al.’s method showed effective KD performance even in an environment with little data. As such, many cases have been reported in which AP’s function to prevent correct classification causes a rather positive effect. We also try to solve the biased sample generation problem by utilizing the AP. Unfortunately, previous AP techniques cannot be used in an environment without label information [10], [12], [13]. Therefore, we propose a new method that induces similar effects to AP even in an environment without label data, and use this method to solve the biased sample generation problem.
FIGURE 1. Distribution of softmax output of the teacher model receiving adversarial samples. For this experiment, WideResNet-40-1 [8] trained with CIFAR-10 was used as a teacher model. The x-axis refers to the class, and is sorted in descending order from the class with the highest average softmax value. Training progressed for a total of 4,000 iterations in this experiment.

FIGURE 2. Conceptual comparison before and after the application of the proposed method. (a) Biased distribution of adversarial samples (b) distribution of adversarial samples after applying the proposed method.

III. METHOD
This section describes the overall flow of the proposed method. First, to solve the problem of biased samples, we propose a method of applying adversarial perturbation where Taylor approximation and convolution of probability distributions are used to give perturbation without label data (Section III.A). Next, a constraint that makes the upper bound of the loss function tight to guarantee the convergence of the loss function to which perturbation has been added is suggested (Section III.B).

A. BIASED SAMPLE GENERATION PROBLEM AND A SOLUTION
A state-of-the-art AL-based C-ZSKD [7] realized AL using Eqs. (2) and (3). AL forces adversarial samples to exist near the decision boundary of the teacher model. So, if the optimal weight for the student model is found by Eq. (3) using adversarial samples, the decision boundary information of the teacher model can be transferred to the student model. On the other hand, \( D_{KL}(T(G(x))|S(G(x))) \) of the right-hand side of Eq. (2) is used as a generator loss function, and is re-defined by Eq. (4).

\[
E_{x \sim Q}[\phi(x)] = -E_{x \sim Q}[D_{KL}(T(x)||S(x))], z \sim N(0, I)
\]  

where \( x = G(z) \), and \( Q \) means an adversarial sample distribution, and \( \phi(x) \) indicates \(-D_{KL}(T(x)||S(x))\). However, when the generator is trained through AL, we can observe that adversarial samples tend to be intensively concentrated on some particular class decision boundaries that are easy to minimize Eq. (4) as shown in Fig. 2 (a). In addition, Fig. 1 shows that adversarial samples are biased to some classes in the early and middle periods of training. Therefore, this phenomenon makes it difficult to deliver accurate decision boundary information of the teacher model to the student model.

This phenomenon can be resolved by moving adversarial samples using the AP in Section 2.2, but the problem is that the AP cannot be used as it is because it requires label data. Therefore, this paper proposes a solution to increase the variance of adversarial sample distribution by using Taylor series approximation and convolution of probability distributions. We will increase the variance of \( Q \) through convolution of \( Q \) and a normal distribution \( A \). However, since \( Q \) is implicitly learned, the exact probability distribution of \( Q \) is unknown. Accordingly, it is impossible to directly calculate the convolution of two probability distributions. Thus, we propose a smart bypass solution using the equivalence relationship in Eq. (5) [15].

\[
E_QE_A[\phi(x + \zeta)] = \int \phi(x) \int q(x - \zeta) \lambda(\zeta) d\zeta dx
= \int \phi(x)(q \ast \lambda)(x) dx
= E_{Q+\lambda}[\phi]
\]
Algorithm 1 Proposed C-ZSKD

1: \( \alpha, \gamma \) and \( \sigma \) are hyperparameters.
2: Let \( \beta = 250, \eta = \text{Learning rate}, N_{\text{total}} = \text{Total iteration number}, N_S = \text{Student model iteration}, T = \text{Teacher model}, S = \text{Student model}, G = \text{generator}, z \sim N(0, I) \).
3: for 1, 2, \ldots, \text{\text{N}}_{\text{total}} \ do
4: \hspace{1cm} \text{Calculate decay factor} \ d_f = \cos(\frac{\text{iter}}{2N_{\text{total}}})
5: \hspace{1cm} \text{Calculate} \ L_1 = -\sigma D_{\text{KL}}(T(x)|S(x)) \ for \ x = G(z)
6: \hspace{1cm} \text{Calculate} \ L_2 = \frac{\partial E}{\partial w} L_1^2
7: \hspace{1cm} \text{Calculate} \ L_3 = \alpha E_x \approx \frac{1}{N} \sum_{i=1}^{N} [\gamma^T \nabla_x L_1(x)]^2, \ \gamma = N(0, \sigma I)
8: \hspace{1cm} \text{Calculate} \ L_G = L_1 + L_2 + L_3
9: \hspace{1cm} \text{Calculate} \ \nabla_{w_G} L_G \text{ for } G
10: \hspace{1cm} \text{Update } w_G \leftarrow w_G - \eta \frac{\partial L_G}{\partial w}
11: \hspace{1cm} \text{for } 1, 2, \ldots, N_S \ do
12: \hspace{2cm} \text{Calculate} \ L_S = D_{\text{KL}}(T(x)|S(x)) \ for \ x = G(z)
13: \hspace{2cm} \text{Calculate} \ L_{\text{AT}} = \beta \sum_{i=1}^{N_S} \left( \frac{j(A_l)}{\|j(A_l)\|_2} - \frac{f(A_l)}{\|f(A_l)\|_2} \right)
14: \hspace{2cm} A_l^T : \text{activation map for } l\text{-th layer of teacher model}
15: \hspace{2cm} A_l^T : \text{activation map for } l\text{-th layer of student model}
16: \hspace{2cm} f(\cdot) : \text{Attention function} \ [14]
17: \hspace{2cm} L_{\text{total}} = L_S + L_{\text{AT}}
18: \hspace{2cm} \text{Calculate} \ \nabla_{w_S} L_{\text{total}} \text{ for } S
19: \hspace{2cm} \text{Update} \ w_S \leftarrow w_S - \eta \frac{\partial L_S}{\partial w}

where \( \Lambda = N(0, \gamma I) \), \( \gamma \) is a hyperparameter, \( \lambda \) and \( q \) are the PDFs of \( \Lambda \) and \( Q \), respectively. Also, \( \xi \sim \Lambda \) and \( x \sim Q \). Eq. (5) can be proved if using the definition of the convolutional integral, the symmetry property of Gaussian distribution, and the communicative property based on the independence of \( x \) and \( \xi \) [15]. Next, if we apply the 2nd-order Taylor series approximation near \( \xi = 0 \) to \( \phi(x + \xi) \) of Eq. (5), Eq. (6) is derived.

\[
\phi(x + \xi) \approx \phi(x) + \left[ \nabla_x \phi(x) \right]^T \xi + \frac{1}{2} \xi^T H(x) \xi \approx \text{Hessian}
\]

(6)

Then, if Expectation with respect to \( \Lambda \) is applied to both sides of Eq. (6), Eq. (7) is obtained through \( \Lambda = N(0, \gamma I) \).

And according to [16], Hessian can be represented through Jacobian \( J \) and approximation error \( R \) as in Eq. (8).

\[
E_\Lambda[\phi(x + \xi)] \approx \phi(x) + \frac{1}{2} \gamma^T Tr(H(x))
\]  
(7)

\[
H(x) = J^T(x)J(x) + R
\]  
(8)

However, referring to [16], the right hand side of Eq. (7) can be represented as Eq. (9) because \( R \) is negligibly small.

\[
\phi(x) + \frac{1}{2} \gamma^T Tr(H(x)) \approx \phi(x) + \frac{1}{2} \gamma^T Tr(J^T(x)J(x))
\]  
(9)

\( Tr(J^T J) \) is squared norm of gradient, so if we apply Expectation with respect to \( Q \) to the right hand side of Eq. (9), Eq. (10) is derived from Eq. (7).

\[
E_{Q*\Lambda}[\phi] \approx E_Q[\phi(x)] + \frac{1}{2} \gamma E_Q[\left[ \nabla_x \phi(x) \right]^2]
\]  
(10)

Finally, Eq. (5) is converted into Eq. (10) through Eqs. (6) to (9), so that the integral operation of Eq. (5) is transformed into an addition of squared norm of gradient and \( E_Q[\phi(x)] \).

Because the adversarial sample distribution is trained through AL and the exact distribution of \( Q \) is unknown, convolution according to Eq. (5) replaces direct convolution. As a result, the adversarial sample distribution becomes \( Q * \Lambda \), so it has a spread distribution compared to the original adversarial sample distribution \( Q \) (see Fig. 2 (b)). Since adversarial samples are distributed near decision boundaries of various classes, the biased sample generation problem can be alleviated.

B. PROPOSED C-ZSKD

Next, we propose a more stable way to learn perturbed samples. The sample distribution estimated by the Taylor approximation shows a clearer decision boundary because it has a higher variance in general. On the other hand, there is a problem that convergence is not guaranteed. To solve this problem, we present a way to tighten the upper bound through an additional constraint. As shown in Eq. (11), the second-order Taylor approximation has an upper bound of \( t(l) \) depending on the loss function \( l \) [17]. Based on this, the approximation constraint is derived as follows. First, using Eq. (6) and Eq. (11), a triangular inequality like Eq. (12) is obtained.

\[
(l(x + y)) - l(x) - \left[ \nabla_x l(x) \right]^T y - \frac{1}{2} \gamma^T H(x) \gamma \leq t(l)
\]  
(11)

\[
\phi(x + \xi) - \phi(x) - \frac{1}{2} \xi^T H(x) \xi \leq \left[ \nabla_x \phi(x) \right]^T \xi
\]  
(12)
follows. Since the biased sample generation problem mainly
occurs in the early and middle of training, the variance of \( \Lambda \) is lowered by applying the decay factor \( d_t \) every iteration. Next, the KL divergence of \( T \) and \( S \) (\( L_1 \)) is calculated. Also, \( L_2 \) is calculated using \( d_t \) and \( L_1 \). After calculating a constraint (\( L_3 \)) for the 2nd-order Taylor approximation, the total loss function of the generator (\( L_G \)) is obtained by adding \( L_1, L_2, \) and \( L_3 \). Finally, the generator weight is updated through the gradient for \( L_G \). As for [7], attention transfer [14] was used as a constraint term for the stability of training, and the distillation loss of [3] was jointly used. The model trained by the proposed algorithm achieves high performance by stably learning adversarial samples with lower bias than the conventional method. The experimental verification is covered in the next section.

IV. EXPERIMENTS

Datasets For fair quantitative evaluation of the proposed method, three datasets were adopted: CIFAR-10 [20], FashionMNIST [21], and CIFAR-100 [22]. Additionally, MNIST [23] was used in the experiment for the analysis of adversarial sample distribution. CIFAR-10 has a training dataset composed of 50,000 RGB images with 10 classes and a size of 32 \( \times \) 32, and a test dataset composed of 10,000 RGB images. FashionMNIST consists of training data composed of 50,000 RGB images with 10 classes and a size of 32 \( \times \) 32, and a test dataset composed of 10,000 RGB images. MNIST has a training dataset consisting of 60,000 gray-scale images.

TABLE 1. Performance comparison of the proposed method, [6] and [7] for the teacher model trained with the CIFAR-10 dataset. Note that the proposed method is improved by an average of 0.52% over [7]. Numerical figures in the last row indicate improvements over [7].

| Method       | [7]  | Ours | [7]  | Ours | [7]  | Ours | [7]  | Ours | [6]  | Ours |
|--------------|------|------|------|------|------|------|------|------|------|------|
| Teacher      | WRN-40-2 | WRN-40-2 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 |
| Student      | 94.7  | 94.7  | 95.2  | 95.2  | 83.0  | 83.7  | 83.7  |
| Teacher Acc. | 83.4  | 83.8  | 90.2  | 90.2  | 81.3  | 81.3  | 81.3  |
| Student Acc. | +0.4  | +0.5  | +0.9  | +0.9  | +0.4  | +0.4  |

TABLE 2. Experimental result on teacher models trained with FashionMNIST. The numerical figures in the last row indicate improvements over [7]. Note that the proposed method improves the performance by 0.78% on average in comparison to [7].

| Method       | [7]  | Ours | [7]  | Ours | [7]  | Ours | [7]  | Ours | [6]  | Ours |
|--------------|------|------|------|------|------|------|------|------|------|------|
| Teacher      | WRN-40-2 | WRN-40-2 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 |
| Student      | 93.6  | 93.6  | 94.9  | 94.9  | 90.8  | 86.1  | 81.6  |
| Teacher Acc. | 82.1  | 84.1  | 86.7  | 87.6  | 89.5  | 89.8  | 79.6  |
| Student Acc. | +2.0  | +0.6  | +0.3  | +0.3  | +0.3  | +0.3  |

TABLE 3. Experimental result on teacher models trained with CIFAR-100. The numerical figures in the last row indicate the performance improvements compared to [7]. Note that the proposed method is 0.58% better than [7].

| Method       | [7]  | Ours | [7]  | Ours | [7]  | Ours | [7]  | Ours | [7]  | Ours |
|--------------|------|------|------|------|------|------|------|------|------|------|
| Teacher      | WRN-40-2 | WRN-40-2 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 | WRN-40-1 |
| Student      | 75.7  | 75.7  | 68.3  | 68.3  | 66.8  | 66.8  |
| Teacher Acc. | 25.6  | 25.9  | 42.6  | 43.0  | 32.4  | 33.3  |
| Student Acc. | +0.3  | +0.4  | +0.4  | +0.4  | +0.9  | +0.9  |

\[
L_G = \sigma E_{\phi}[\phi(x)] + \frac{1}{2} \nu E_{\phi}[(\nabla_{\phi} \phi(x))^2] + \alpha E_{\phi}[(\nabla_{\phi} \phi(x))^T \zeta] \tag{14}
\]

where \( \alpha \) and \( \sigma \) are the hyperparameters. Here, the hyperparameters are determined according to the intrinsic dimensionality of latent space [18], [19] of the dataset.
FIGURE 3. (a) Distribution of MNIST test dataset in the embedding space. (b) to (d) mean the distributions in the embedding space of the generated adversarial samples when the generator and student model are trained by the teacher model trained with MNIST. The tops of (b) to (d) are the distribution of adversarial samples generated by [7]. The bottoms of (b) to (d) are the distribution of adversarial samples generated by the proposed method. The bar on the right in each figure represents density.

Implementation Detail

We adopted WideResNet-40-2 [8] as a teacher model. When training the generator and student model of [7], as well as the proposed method, AdamW [24] was used as an optimizer. In AlexNet and LeNet, the initial learning rates of the generator and student model were set to $5 \times 10^{-3}$ and $10^{-3}$, respectively. The initial learning rate of the generator and student model in the other networks was set to $2 \times 10^{-3}$. And $\beta_1$ and $\beta_2$ were set to 0.9 and 0.999, respectively, and the weight decay was $10^{-2}$. The learning rate was adjusted to decrease gradually every iteration by referring to the cosine annealing proposed by [25]. The batch size was set to 512, and the number of iterations of AL was 16,000 times. $N_S$, which determines the AL ratio of the generator and student model, was set to 5. When training the student model through the teacher model trained with MNIST, the batch size was set to 64, and a total of 4,000 iterations were trained.

Note that the generator structures when training CIFAR-10 and CIFAR-100 are different from those when training MNIST and FashionMNIST. In the proposed C-ZSKD, the purpose of the generator is to increase the logit distribution difference between the teacher model and the student model as shown in Eq. (2). MNIST and FashionMNIST are composed of relatively monotonous images compared to CIFAR-10 and CIFAR-100. A dataset composed of monotonous images has a relatively low dimensional latent space according to the manifold hypothesis. So, if a generator learns a teacher model trained with a monotonous dataset, the generator is likely to cause mode collapse. Therefore, when training through a teacher model trained with MNIST and FashionMNIST, we reduced the parameter size of the generator and added an average pooling layer to the generator for suppressing this phenomenon effectively. And $\sigma$ was set to $10^{-3}$ which is very small compared to CIFAR-10 and CIFAR-100 for improving the training stability.

A. QUANTITATIVE RESULTS

In order to quantitatively evaluate the proposed method, the classification accuracy for three test datasets, i.e., CIFAR-10, CIFAR-100, and FashionMNIST, was measured. The hyperparameters $\alpha$, $\gamma$, and $\sigma$ in Eq. (14) were determined according to the dataset used for the teacher model. In detail, $\gamma$ was determined proportionally to the inter-class variability of the dataset for mitigating the biased sample generation problem, and $\alpha$ was determined experimentally considering the stability of training. As a result, $\alpha = 5$, $\gamma = 2.5$, $\sigma = 1$ for CIFAR-10. In CIFAR-100, $\alpha = 20$, $\gamma = 4$, $\sigma = 1$. Lastly, for FashionMNIST, $\alpha = 1$, $\gamma = 2$, $\sigma = 10^{-3}$. Here, the hyperparameters were set empirically.
TABLE 4. Effects of hyperparameters on the overall performance of the proposed method. The 'baseline' stands for the result from Table 1, and 'difference' indicates the gap from the 'baseline.'

|            | Baseline | $\alpha = 3$ | $\alpha = 7$ | $\gamma = 2$ | $\gamma = 3$ | $\sigma = 0.8$ | $\sigma = 1.2$ |
|------------|----------|---------------|---------------|---------------|---------------|----------------|----------------|
| Student Acc | 90.2     | 89.9          | 89.6          | 89.2          | 89.3          | 90.1           | 89.8           |
| Difference  | -0.3     | -0.6          | -1.0          | -0.9          | -0.1          | -0.4           | -0.4           |

are somewhat sensitive, they can cause performance degradation if they are not appropriately set. This is verified in the ablation study. Using the same seeds, a total of three trainings were performed on all teacher-student pairs, and the average accuracy was given as student accuracy.

Tables 1 to 3 compare the student accuracy of the proposed method and the conventional C-ZSKD techniques for CIFAR-10, FashionMNIST, and CIFAR-100. First, the proposed method provides 0.52%, 0.78%, and 0.58% better accuracy on average than the conventional state-of-the-art technique [7] for the three datasets, respectively, and shows up to 0.86%, 2.06%, and 0.90%. We can observe that the performance improvement of the proposed method is noticeable as the channel width of the student model is smaller regardless of the dataset. In other words, the smaller the number of channels in the student model, the more serious the biased sample generation problem. As a result, the biased sample generation problem was significantly alleviated by the proposed method.

Next, take a look at the difference in performance between the teacher model and the student model. Comparing Table 1 and Table 3, we can observe that as the number of classes increases, the performance difference between the two models increases. When training with CIFAR-100 (see Table 3), the difference in performance between the two models amounts to 35.56% on average. Such a big difference means that the adversarial samples do not accurately transfer the decision boundary of the teacher model trained with CIFAR-100. In other words, it is more difficult to transfer the exact decision boundary of a teacher model trained with a dataset with many classes. We can analyze that this phenomenon occurs because the adversarial samples may omit decision boundary information of some classes. In the end, the more complex the decision boundary, as in training CIFAR-100, the more serious the biased sample generation problem. However, note that this problem is mitigated when the adversarial samples generated by the proposed method are applied to knowledge distillation. That is, as shown in Table 3, the proposed method delivers the decision boundary of the teacher model to the student model more accurately.

B. ANALYSIS FOR EFFECTIVE ADVERSARIAL SAMPLES

This section analyzes the characteristics of adversarial samples effective for AL-based C-ZSKD. To this end, the distribution in the embedding space of the network pre-trained with the MNIST dataset was visualized. Fig. 3 (a) shows the distribution of the MNIST test dataset embedded by the modified WideResNet trained with MNIST training dataset. The farther a sample is from the origin (0,0), the more certain the class is. Fig. 3 (b), (c), and (d) describe the distributions of adversarial samples represented in the same coordinate system as in Fig. 3 (a) for several iterations. We can observe that adversarial samples generated by the proposed method are more concentrated on the origin. On the other hand, adversarial samples generated by [7] show a relatively biased distribution. This indicates that the proposed method produces more useful adversarial samples than [7]. Therefore, it was experimentally proved that the proposed method greatly mitigates the biased sample generation problem, which is an inherent problem of AL-based C-ZSKD.

Based on the experiments of Sections 4.1 and 4.2, we can interpret that adversarial samples have ambiguity because they are close to the origin of Fig. 3 (a). This characteristic contributes a lot to transfer decision boundary information of as many diverse classes as possible.

C. ABLATION STUDY

This section describes the ablation study for each hyperparameter of the proposed method. The dataset used in this experiment was CIFAR-10, and WRN-40-2 and WRN-16-2 were used as teacher and student networks, respectively. Table 4 shows that performance deteriorates when each hyperparameter deviates from the optimal value. This indicates that, as mentioned above, the proposed method has a somewhat sensitive property to hyperparameters. In particular, since the sensitivity to $\gamma$ is high, careful adjustment is required. It must be one weakness that the proposed method is sensitive to hyperparameter settings. Nevertheless, the proposed method is valuable because it can provide sufficiently higher performance than the existing C-ZSKD.

V. CONCLUSION

This paper aims to solve the biased sample generation problem of the complete zero-shot knowledge distillation (C-ZSKD) based on adversarial learning (AL). Inspired by the conventional C-ZSKD based on adversarial learning, we have devised a novel method that can have a similar effect even in an environment without label data. The proposed method increases the variance of the adversarial sample distribution by using the squared gradient norm of the generator loss function. As a result, the adversarial sample distribution is widened, so the student model is trained by receiving more accurate decision boundary information from the teacher model. The experiments on various datasets showed that the student model trained by the proposed method could provide a high performance improvement of up to 2.06%. Additionally, by analyzing the distribution of adversarial samples on the embedding space, the characteristic of the most effective
adversarial samples for AL-based C-ZSKD is qualitatively demonstrated. In the future, we will expand the AL-based C-ZSKD study in the direction of generating adversarial samples with high entropy in the embedding space.

REFERENCES

[1] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[3] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531. [Online]. Available: http://arxiv.org/abs/1503.02531

[4] R. G. Lopes, S. Fenu, and T. Starner, “Data-free knowledge distillation for deep neural networks,” 2017, arXiv:1710.07535. [Online]. Available: http://arxiv.org/abs/1710.07535

[5] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, “High accuracy and high fidelity extraction of neural networks,” in Proc. 29th USENIX Secur. Symp., 2020, pp. 1345–1362.

[6] G. K. Nayak, K. R. Mopuri, V. Shaj, V. B. Radhakrishnan, and A. Chakraborty, “Zero-shot knowledge distillation in deep networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 4743–4751.

[7] P. Micelli and A. J. Storkey, “Zero-shot knowledge transfer via adversarial belief matching,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 9547–9557.

[8] S. Zagoruyko and N. Komodakis, “Wide residual networks,” 2016, arXiv:1605.07146. [Online]. Available: http://arxiv.org/abs/1605.07146

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, arXiv:1412.6572. [Online]. Available: http://arxiv.org/abs/1412.6572

[10] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1765–1773.

[11] B. Heo, M. Lee, S. Yun, and J. Y. Choi, “Knowledge distillation with adversarial samples supporting decision boundary,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 3771–3778.

[12] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, “Generative adversarial perturbations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4422–4431.

[13] M. M. Naseer, S. H. Khan, M. H. Khan, F. S. Khan, and F. Porikli, “Cross-domain transferability of adversarial perturbations,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 12885–12895.

[14] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” 2016, arXiv:1612.03928. [Online]. Available: http://arxiv.org/abs/1612.03928

[15] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, “Stabilizing training of generative adversarial networks through adversarial layer transformation,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 2018–2028.

[16] G. Pratt, C. Shin, and G. J. Hick, “Gauss–Newton and full Newton methods in frequency-space seismic waveform inversion,” Geophys. J. Int., vol. 133, no. 2, pp. 341–362, May 1998.

[17] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, “Robustness via curvature regularization, and vice versa,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9078–9086.

[18] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” 2017, arXiv:1711.01558. [Online]. Available: http://arxiv.org/abs/1711.01558

[19] P. K. Rubenstein, B. Schoelkopf, and I. Tolstikhin, “On the latent space of Wasserstein auto-encoders,” 2018, arXiv:1802.03761. [Online]. Available: http://arxiv.org/abs/1802.03761

[20] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Tech. Rep., 2009.

[21] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017, arXiv:1708.07747. [Online]. Available: https://arxiv.org/abs/1708.07747

[22] A. Krizhevsky, V. Nair, and G. Hinton, CIFAR-100 (Canadian Institute for Advanced Research). Accessed: Nov. 24, 2020. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[24] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, arXiv:1711.05101. [Online]. Available: http://arxiv.org/abs/1711.05101

[25] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” 2016, arXiv:1608.03983. [Online]. Available: http://arxiv.org/abs/1608.03983

KANG IL LEE received the B.S. and M.S. degrees in electronic engineering from Inha University, Incheon, South Korea, in 2018 and 2020, respectively. His research interests include computer vision and machine learning.

SEUNGHYUN LEE (Associate Member, IEEE) received the B.S. degree in electronic engineering from Inha University, Incheon, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electronic engineering. His research interests include computer vision and machine learning.

BYUNG CHEOL SONG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1994, 1996, and 2001, respectively. From 2001 to 2008, he was a Senior Engineer with the Digital Media Research and Development Center, Samsung Electronics Company Ltd., Suwon, South Korea. In 2008, he joined the Department of Electronic Engineering, Inha University, Incheon, South Korea, where he is currently a Professor. His research interests include image processing and computer vision.