Distillation for High-Quality Knowledge Extraction via Explainable Oracle Approach

MyungHak Lee
Wooseong Cho
Sungik Kim
Jinkyu Kim
Jaekoo Lee

1 College of Computer Science
Kookmin University
Seoul, Korea

2 Department of Computer Science and Engineering
Korea University,
Seoul, Korea

Abstract

Recent successes suggest that knowledge distillation techniques can usefully transfer knowledge between deep neural networks as compression and acceleration techniques, e.g., effectively and reliably compress a large teacher model into a smaller student model with limited resources. However, knowledge distillation performance is degraded when the model compression rate becomes excessively high due to the size of the teacher model. To address this, we advocate for improving the teacher-to-student knowledge transfer by identifying and reinforcing input-level signals of substantial contributions for a final verdict, e.g., signals for a long trunk of elephants are strengthened and transferred to the student model. To this end, we adopt gradient-based explainable AI techniques for extracting output-relevant input-level features. Then, we strengthen and transfer these signals to improve the knowledge distillation performance. Our experiments on public datasets (i.e., CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet) show that our method clearly outperforms existing knowledge distillation approaches, especially in the case of using a small teacher model. Our code is available at https://github.com/myunghakLee/Distillation-for-High-Quality-Knowledge-Extraction.

1 Introduction

The objective of knowledge distillation (KD) is to facilitate the transfer of knowledge from one model (a teacher) to another model (a student) that is typically simpler without loss of validity. As shown in the Figure 1. (a), most previous KD methods use two types of knowledge (i.e., feature-based and response-based knowledge) extracted by a pre-trained and frozen teacher model to transfer knowledge into the student model [18, 42]. In general, the student model which leverages knowledge leads to enhanced performance relative to training solely on actual labels [22]. Moreover, KD serves as a regularizer, progressively employing fewer basis functions for the iterative learning of a model [8].

In a deep model, each layer learns different levels of feature representation with increasing abstraction [21, 18]. Therefore, most KD methods use a teacher model that is equal to or
Figure 1: (a) An overview of conventional knowledge distillation techniques that use two main approaches: feature distillation and response distillation. (b) Classification accuracy of the student model degrades (with existing knowledge distillation techniques, such as DKD [60], ATT [59], FitNet [42], KD [18], and SAD [21]) as the teacher-student model capacity gap increases.

larger than the student model, in order to utilize knowledge of a higher level of abstraction. However, when a disproportionately large teacher model is utilized, the student model may fail to appropriately receive the knowledge from the teacher model due to the significant capacity gap between the models [13, 27, 32]. Additionally, there are other distillation issues with establishing links of the knowledge between the teacher and the student models [20, 21]. For these reasons, existing KD methods cannot guarantee performance enhancement of the student model in response to an increase in depth of the teacher model (refer to Figure 1. (b)).

Recent work tackles effective knowledge transfer despite the huge capacity gap between the teacher and student models [21, 23, 27]. However, it is still challenging to achieve a successful KD that can help mitigate the impact of the capacity gap. To address these challenges, we propose a novel KD method that extracts high-quality knowledge by reinforcing data via explainable AI (XAI) technique.

In particular, we generate relevance-reinforced data using XAI and adversarial example techniques. This enables the extraction of high-quality knowledge even from the limited teacher model. Our Oracle approach, which uses XAI techniques to reinforce input pixels that help reduce task loss and diminish input pixels that hinder it, can be easily extended to other distillation schemes. We summarize our main contributions as follows:

• We propose a novel knowledge distillation method that can extract high-quality knowledge via explainable AI and adversarial example.

• We effectively show the benefit of our method on public datasets: CIFAR-10, CIFAR-100 [24], Tiny-ImageNet, and ImageNet [8].

• Our method quantitatively and qualitatively outperforms alternative knowledge distillation methods.

2 Related Work

2.1 Knowledge Distillation (KD)

Knowledge distillation (KD) is a widely used technique that trains a student model under the supervision of a pre-trained teacher model [18, 42]. KD has been successfully applied to several learning tasks such as image classification [18, 42], object detection [5, 26], and image segmentation [29, 56]. Recent work can be broadly divided into two approaches concerning
extraction and distillation of the knowledge from teacher to student without a significant performance drop. Several approaches address how to maximize the benefit of the teacher model by reducing the capacity gap experienced by the student model \([6, 23, 31, 36]\). In addition, Sau et al. \([44]\) propose a noise-based regularizer from multi-teacher models, and Wang et al. \([57]\) discover the interplay between KD and data augmentation. Zhao et al. \([60]\) reformulate KD loss into target class and non-target classes KD and Jang et al. \([20]\) introduce a meta-learning approach capable of automatically determining the knowledge to transfer from the one model to where in another. A few recent works \([21, 40]\) use attention mechanism to match feature-level in KD. ResKD \([27]\) uses the capacity gap between teacher and student models as guidance to train a significantly more lightweight student model. In KCD \([25]\), the knowledge value on each sample is dynamically estimated by EM algorithm to distill a compact knowledge set from the teacher model, thereby guiding student training.

### 2.2 Adversarial Example

An adversarial example (attack) is an instance with subtle, deliberate perturbations in features, compelling a learning model \(f(\cdot)\) to make incorrect predictions \([14, 52]\). The equation is defined as follows:

\[
f(x) \neq f(x + \varepsilon), \quad \|\varepsilon\|_2 < \eta, \quad \varepsilon = \gamma \text{sign}(\nabla_x f(\theta, x, y))
\]

where \(\eta\) is a perturbation constraint, \(\theta\) represents the parameters of the model, and \(\gamma\) adjusts the intensity of the perturbation. Schmidt et al. \([46]\) introduce the generalization of adversarially robust learning by investigating the sample complexity required to achieve robustness against adversarial examples. Farnia et al. \([11]\) provide bounds on the generalization error for deep neural networks trained under several adversarial attack schemes. Miyato et al. \([34]\) extend adversarial learning and its examples to natural language domains.

### 2.3 Explainable AI (XAI)

Explainable AI (XAI) aims to AI decision-making. Recent AI models, akin to the black box, have challenges in discerning the rationale behind their results. This hinders not only performance enhancement but also advancements in building trustworthiness in AI. From DARPA \([54]\), the field of XAI is broadly presented into two categories; 1) to produce more interpretable models, ensuring a high level of performance and 2) to enable humans to comprehend and appropriately trust AI. In deep learning, it provides explainability through visualization of saliency \([4, 22, 38, 47, 61]\) and relevance signals \([1, 49]\). Additionally, there are works employing XAI techniques to refine data \([0, 51]\), feature \([0, 63, 69, 71, 72, 73]\), gradient \([64, 72]\), and loss \([0, 51, 63, 72, 82, 83, 84]\) with the aim of improving model performance. Inspired by the aforementioned works, we propose a novel methodology to enhance KD performance by reinforcing input with the perturbation generated by XAI technique.

### 3 Method

To extract high-quality knowledge, previous studies usually increase the size of the teacher model. However, when employing an excessively large teacher model, the performance of the student model may decrease because of the capacity gap. Due to this issue, increasing the size of the teacher model cannot improve the student model’s performance beyond a certain
threshold. In other words, there are limitations in the quality of knowledge extracted from the teacher model and transferred to the student model. To address these issues, we propose an Oracle teacher model that extracts high-quality knowledge through the use of reinforced data, generated by XAI.

3.1 Step A: Generating Relevance-Reinforced Inputs

As shown in Figure 2 (left), we first identify pixel-level contributions for the final verdict, i.e., which parts of an input image largely (or negligibly) contribute to the model to draw its output. Determining each pixel’s contribution has widely been explored as a tool to build explainable (or interpretable) models in previous works. Following these works, we want to quantify input-level contribution by computing gradients of the task loss function. Formally, we compute the task loss $L_{\text{task}}$ given the ground truth $y$ and the predicted output $f_\theta(x)$ with a model $f_\theta$ parameterized by $\theta$. By applying the standard backpropagation method, we compute gradients $g = \frac{\partial f(x)}{\partial x}$ to determine the amount of relevance scores of an input $x$. Given this gradient $g$, we modify the input by pushing it toward the negative direction of gradients, i.e., we obtain a modified input where parts of high relevance scores are reinforced, producing much more confident decisions when it is used as the input itself. The equation is following:

$$x^* = x - \gamma |x| \odot \frac{\partial f(x)}{\partial x} \quad (2)$$

where $\odot$ represents the element-wise multiplication and $\gamma$ is a hyperparameter to control the strength of this modification.

As reported in Table 1, such a relevance-reinforced input $x^*$ provides a dramatic performance boost in all backbone types (compare the last two columns). These results may be intuitive as we use the ground truth to drive the model to draw the correct output by modifying inputs (i.e., the Oracle model can achieve dramatic performance improvements due to its prior knowledge of the ground truth); thus, this may not be useful for inference in common real-world scenarios. However, this may be useful for the knowledge distillation task where

Figure 2: An overview of our proposed knowledge distillation method, which consists of two main steps: (A) Generating Relevance-Reinforced Inputs and (B) Transfer Knowledge via Oracle Teacher Model. In Step A, we generate reinforced data $x^*$ where input pixels that make the model correctly classify are reinforced. Further, in Step B, the generated reinforced data is then used to extract the teacher model’s responses for the classification task, transferring them into the student model.
identifying which parts of an image are focused on by a teacher model. We, therefore, want to leverage the relevance-reinforced input as a key component to transfer knowledge to the student model.

### 3.2 Step B: Transfer Knowledge via Oracle Teacher Model

Following the standard response-based knowledge distillation technique [18], we first compute the soft predictions \( P_T \) from the last layer of our teacher model, i.e., a probability distribution over different categories with a softmax layer: \( P_T = \text{Softmax}(o_T/\tau) \), where \( o_T \) is the final output feature from the teacher model. We also use the temperature \( \tau \) to prevent overconfidence issues during training. Further, we distill the teacher model’s response knowledge to the student model by minimizing the KL divergence of soft predictions as follows:

\[
L_{\text{distill}} = D_{KL}(P_T || P_S)
\]

where \( P_S = \text{Softmax}(o_S/\tau) \) is the soft prediction from the last layer of the student model.

**Loss Function.** We train our model end-to-end by minimizing the following loss \( \mathcal{L} \):

\[
\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{distill}}
\]

where \( \mathcal{L}_{\text{cls}} = H(P_S, y) \) is the cross-entropy loss for the classification task from the student model. We use hyperparameter \( \alpha \) to control the strength of the distillation loss term.

### 4 Experiments

Through various experiments, 1) we validate that our Oracle teacher model made by using XAI extracts high-quality knowledge and 2) conduct a comparative analysis between our Oracle teacher model and SOTA knowledge distillation methods to demonstrate that the proposed method significantly enhances the performance of the student model without increasing the size of the teacher model.

### 4.1 Is the Knowledge Obtained from Our Oracle Teacher Model Good Enough?

To validate whether the extracted knowledge from the Oracle teacher model can effectively train the student model, we conducted the following four experiments: (i) We measure the Expected Calibration Error (ECE) [15] of the Oracle teacher model, aiming for measuring the quality of the response knowledge of train dataset as presented in Table 2. (ii) We provide visualization results using t-SNE for the output representation of the Oracle teacher model. (iii) We measure the silhouette score of the output representation of the Oracle teacher
Table 2: The number of parameters and ECE scores for ResNet-based variant models on ImageNet dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th># of Param</th>
<th>Scratch model</th>
<th>Oracle model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>11.2M</td>
<td>0.0327</td>
<td>0.0311</td>
</tr>
<tr>
<td>ResNet34</td>
<td>21.3M</td>
<td>0.0313</td>
<td>0.0247</td>
</tr>
<tr>
<td>ResNet50</td>
<td>23.7M</td>
<td>0.0284</td>
<td>0.0216</td>
</tr>
<tr>
<td>ResNet101</td>
<td>42.7M</td>
<td>0.0237</td>
<td>0.0164</td>
</tr>
</tbody>
</table>

model [6, 34, 31]. Lastly, (iv) we measure the entropy whether the knowledge from the teacher model can effectively be compressed to alleviate the capacity gap issue.

Expected Calibration Error (ECE) Analysis. Expected Calibration Error is a metric that approximates how well the confidence scores $conf(\cdot)$, obtained by applying softmax to the logits of a classification model, align with the actual probability $acc(\cdot)$ of a correct prediction. It is measured as follows:

$$ECE = \sum_{b=1}^{B} \frac{|b|}{N} |acc(b) - conf(b)|$$  \hspace{1cm} (5)

where $conf(b) = \frac{1}{|b|} \sum_{j \in b} p_j$ and $acc(b) = \frac{1}{|b|} \sum_{j \in b} 1(p_j = y_j)$ in bin $b$. $N$ is the total number of data samples.

The ECE, which quantifies the difference between $conf(\cdot)$ and $acc(\cdot)$, is commonly used as a metric to evaluate the level of overconfidence of the model. In recent deep neural networks, it is often observed that increasing the model’s capacity improves accuracy but can lead to higher ECE due to overconfidence [15]. If a model exhibits not only high accuracy but also low ECE, it can be considered as extracting good knowledge for classification. Therefore, we verify whether our Oracle teacher model extracts high-quality knowledge that strikes a balance between accuracy and confidence.

Our Oracle teacher model demonstrated high performance in both accuracy and ECE metrics. In Table 1, it consistently shows better (higher) accuracy that is up to 1.4 times (ResNet18), and in Table 2, it consistently shows better (lower) ECE that is up to 1.4 times (ResNet101). This result arises from the fact that reinforced data serves to increase the confidence scores for the target class while decreasing those for the non-target classes. Consequently, we confirm that the response knowledge extracted from our Oracle teacher model is superior to that of conventional scratch models.

t-SNE Analysis. t-SNE is used for computing pairwise similarities of classes in the latent space and visualizing in a low dimensional space [35]. It is generally observed that semantically similar inputs tend to evoke similar activation patterns in a trained model [35, 33]. If knowledge is extracted that enhanced the distinct representation among similar inputs, it can lead to improved performance of classification. Consequently, we expect to observe well-distinguishable clusters among similar inputs. In the well-known CIFAR-100 dataset [24], each image comes with a fine label as sub-classes and a coarse label as super-classes. We use these super-classes as similar inputs and evaluate the quality of extracted knowledge by the models. In Figure 3, we visually confirm that clustering is more distinct in our Oracle teacher model ($\gamma = 0.5, 1.0$) compared to clustering in the scratch model ($\gamma = 0$).

Silhouette Score. Silhouette coefficient is a metric that validates consistency within the cluster. It is calculated for each data as shown in Eq. 6, and the value for the entire dataset is derived by taking the average. In this paper, we refer to this averaged value as the sil-
Figure 3: Visualizations by t-SNE [55] for output representation of the our Oracle teacher model (ResNet20) on CIFAR-100 dataset with varying $\gamma$ in Eq. 2. For better understanding, we provide sample images and denote color coding points according to their superclass.

Silhouette score, with values closer to 1 indicating better performance. $D_{ic}(k)$ indicates the intra-cluster distance, the mean distance between $k$ and all other data points in the same cluster, $D_{ic}(k) = \frac{1}{|S_K|} \sum_{j \in S_K, k \neq j} d(k, j)$, and $D_{nc}(k)$ indicates the nearest-cluster distance, the smallest mean distance of $k$ to all data in any other cluster, $D_{nc} = \min_{L \neq K} \frac{1}{|S_L|} \sum_{j \in S_L} d(k, j)$.

$$\text{Sil}(k) = \frac{D_{nc}(k) - D_{ic}(k)}{\max(D_{nc}(k), D_{ic}(k))}$$

where $k$ is data in cluster $K$, $S_K$ is set of cluster $K$, and $d(\cdot)$ is a distance function.

As depicted in Figure 4, variations in the silhouette score are measured by varying $\gamma$ in Eq. 2. As can be seen, silhouette scores for both super-class and sub-class increase as $\gamma$ grows up to a certain threshold. This implies that our Oracle teacher model ($\gamma = 0.5, 1.0$) extracts superior knowledge compared to the scratch model ($\gamma = 0$) in terms of the silhouette score. However, if $\gamma$ is set too high, we observe a drop in the silhouette score. Therefore, we confirm that the optimal $\gamma$ can be explored through the silhouette score.

**Entropy.** In our proposed approach, we elevate the confidence score for the target class while attenuating it for non-target classes. This action effectively reduces the entropy (i.e., reduces the amount of information) of our knowledge, as outlined in Table 3. Nonetheless, empirical evidence from our experiments suggests, that information among similar classes is

Figure 4: Variation of silhouette score and top-1 accuracy with varying $\gamma$ in Eq. 2. The best and maximum scores are indicated by a circle.
Table 3: Comparison of entropy between confidence score about original data \((x)\) and Refined data \((x^*)\) on two datasets: CIFAR-10 and CIFAR-100.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data ((x))</td>
<td>5.39</td>
<td>5.30</td>
</tr>
<tr>
<td>Reinforced data ((x^*))</td>
<td>5.01</td>
<td>4.87</td>
</tr>
</tbody>
</table>

Table 4: Classification accuracy of the student model with different knowledge distillation methods applied. We use CIFAR-10 and CIFAR-100 datasets. Note that bold represents the best score, and the underlined scores represent the second best. Accuracy of the teacher model is also shown in parentheses.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR-10 dataset</th>
<th>CIFAR-100 dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKD [60]</td>
<td>90.27</td>
<td>91.07</td>
</tr>
<tr>
<td>ATT [59]</td>
<td>91.11</td>
<td>91.88</td>
</tr>
<tr>
<td>FitNet [42]</td>
<td>92.17</td>
<td>93.06</td>
</tr>
<tr>
<td>KD [18]</td>
<td>92.57</td>
<td>93.58</td>
</tr>
<tr>
<td>SAD [21]</td>
<td>92.71</td>
<td>93.61</td>
</tr>
<tr>
<td>Oracle (Ours)</td>
<td>92.94</td>
<td>93.77</td>
</tr>
</tbody>
</table>

preserved (refer to Figure 3 and Table 2). Consequently, our method preserves valid information while diminishing the total amount of information. Therefore, our method can alleviate the capacity gap issue because it only needs to transfer a smaller amount of information to the student model.

4.2 Knowledge Distillation Performance Comparison

We further compare the knowledge distillation performance with other existing SOTA models, including DKD [60], ATT [59], FitNet [42], KD [18], and SAD [21]. Following the standard setting for the knowledge distillation task, we conduct two scenarios: (i) The teacher and student models have the same backbone (i.e., self-distillation), and (ii) The teacher and student models have and do not have different backbones. Our experiment is based mainly on ResNet-based backbones [17] with different numbers of layers, i.e., 18, 20, 32, 34, 56, and 110, while we use the following four publicly available image classification datasets: CIFAR-10, CIFAR-100 [24], ImageNet, and Tiny-ImageNet [8].

Self-distillation Performance. We first observe in Table 4 that our model clearly outperforms the other knowledge distillation methods on two datasets (CIFAR-10 and CIFAR-100) in cases where the teacher and student models have the same backbones (which could be called self-distillation [35]). This performance gain is consistently observed for all types of backbones.

Knowledge Distillation Performance. Further, we also compare the knowledge distillation performance in cases where the teacher and student models have and do not have the same backbone. We compare the student model’s classification accuracy on CIFAR-100 dataset with the other KD methods, including DKD, ATT, FitNet, KD, and SAD. We test 16 different pairs of ResNet-based backbones. As depicted in Figure 5, ours demonstrates a superior performance across overall different pairs for each student model. Interestingly, despite setting the teacher and student models to be the same, ours outperformed all other model pairs using the same student model in other KD methods. This demonstrates the effectiveness of our method in extracting and transferring high-quality knowledge, even with a smaller teacher...
resNet20 (Student)  |  resNet32 (Student)  |  resNet56 (Student)  |  resNet110 (Student)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Model</td>
<td>Teacher Model</td>
<td>Teacher Model</td>
<td>Teacher Model</td>
</tr>
<tr>
<td>resNet20</td>
<td>resNet32</td>
<td>resNet56</td>
<td>resNet110</td>
</tr>
</tbody>
</table>

Figure 5: Comparison of accuracy of our Oracle and SOTA knowledge distillation methods on CIFAR-100 dataset. X-axis represents a different teacher model and Y-axis represents the accuracy. We denote the best performing method for each student model with a star symbol.

Table 5: Classification accuracy on Tiny-ImageNet and ImageNet datasets. Note that bold represents the best score, and the underlined scores represent the second best. Accuracy of the scratch model is also shown in parentheses.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tiny-ImageNet</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet18 (59.83)</td>
<td>ResNet34 (61.50)</td>
<td>ResNet34 (73.31)</td>
</tr>
<tr>
<td>Student model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet18 (59.83)</td>
<td>ResNet34 (61.50)</td>
<td>ResNet34 (73.31)</td>
</tr>
<tr>
<td>DKD [60]</td>
<td>64.15</td>
<td>66.89</td>
</tr>
<tr>
<td>ATT [59]</td>
<td>63.58</td>
<td>65.04</td>
</tr>
<tr>
<td>FitNet [42]</td>
<td>65.56</td>
<td>67.63</td>
</tr>
<tr>
<td>KD [18]</td>
<td>65.13</td>
<td>67.43</td>
</tr>
<tr>
<td>SAD [21]</td>
<td>65.78</td>
<td>67.21</td>
</tr>
<tr>
<td>Oracle (Ours)</td>
<td><strong>66.25</strong></td>
<td><strong>67.99</strong></td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, we propose using gradient-based explainable AI techniques to improve the model performance and compression effect of knowledge distillation techniques effectively, reducing the commonly observed degradation issue of the student model given a large teacher-student model capacity gap. Show that our method clearly outperforms existing knowledge distillation approaches, when we set the teacher and student models to be the same, ours performs better than all others using the same student model. Plus, our analysis demonstrates the validity and usefulness of that reinforced teacher-to-student knowledge with Expected Calibration Error (ECE), t-SNE, Silhouette Score and Entropy.

Acknowledgement. This research was supported by the National Research Foundation(NRF) grant (No.RS-2023-00212484) and Institute of Information & Communications Technology Planning & Evaluation(IITP) grant (No.RS-2022-00167194) funded by the Korea government(MSIT). This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(2022-0-00043, 20%), and the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2023-2020-0-01819, 10%) supervised by the IITP.
References


