Distillation for High-Quality Knowledge Extraction via Explainable Oracle Approach

MyungHak Lee[1]
Wooseong Cho[1]
Sungsik Kim[1]
Jinkyu Kim[*2]
jinkyukim@korea.ac.kr
Jaekoo Lee[*1]
jaekoo@kookmin.ac.kr

[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 [35]. Moreover, KD serves as a regularizer, progressively employing fewer basis functions for the iterative learning of a model [35].

In a deep model, each layer learns different levels of feature representation with increasing abstraction [3, 37]. Therefore, most KD methods use a teacher model that is equal to or

* Corresponding authors

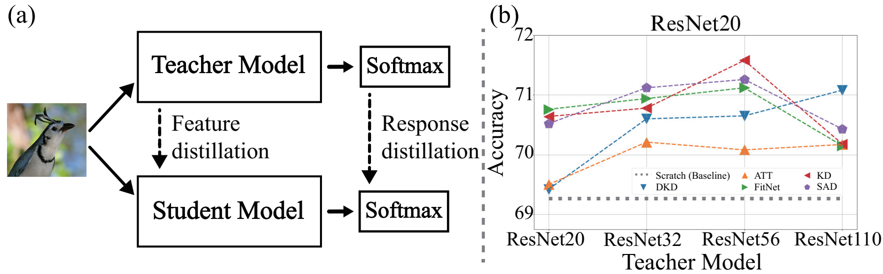

Figure 1: (a) An overview of conventional knowledge distillation techniques that use two main approaches: feature distillation and response distillation. (b) Classification accuracy of the student model degrades (with existing knowledge distillation techniques, such as DKD [60], ATT [59], FitNet [42], 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:

$$f(\mathbf{x}) \neq f(\mathbf{x}+\varepsilon), \ \|\varepsilon\|_2 < \eta \ , \qquad \varepsilon = \gamma \,\mathrm{sign}\left(\nabla_{\mathbf{x}} f(\theta, \mathbf{x}, y)\right) \tag{1}$$

where η is a perturbation constraint, θ represents the parameters of the model, and γ 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 [2, 30], feature [12, 33, 45, 51, 62, 63], graident [16, 39], and loss [9, 10, 19, 28, 41, 48, 58] 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

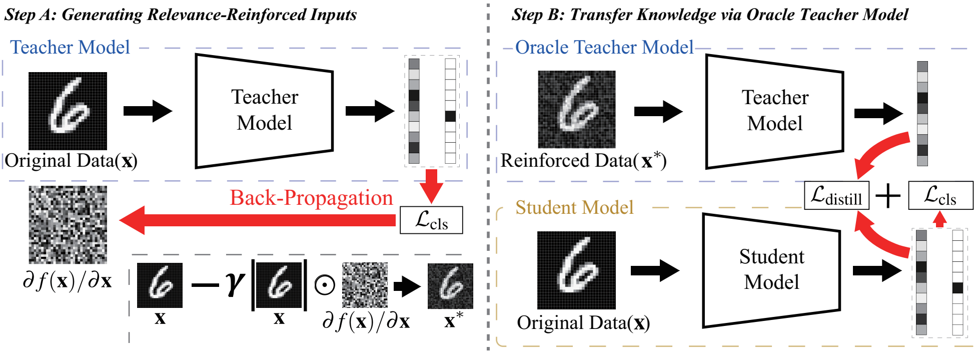

Figure 2: An overview of our proposed knowledge distillation method, which consists of two main steps: (A) Generating Relevance-Reinforced Inputs and (B) Transfer Knowledge via Oracle Teacher Model. In Step A, we generate reinforced data $\mathbf{x}^*$ where input pixels that make the model correctly 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.

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 $\mathcal{L}_{\text{task}}$ given the ground truth y and the predicted output $f_\theta(\mathbf{x})$ with a model f_θ parameterized by θ. By applying the standard backpropagation method, we compute gradients $\mathbf{g} = \partial f(\mathbf{x})/\partial \mathbf{x}$ to determine the amount of relevance scores of an input $\mathbf{x}$. Given this gradient $\mathbf{g}$, 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:

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

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

As reported in Table 1, such a relevance-reinforced input $\mathbf{x}^*$ 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

Table 1: Comparison of accuracy between scratch models trained by original data ($\mathbf{x}$) and Oracle models trained by reinforced data ($\mathbf{x}^*$) on two datasets: CIFAR-100 and ImageNet.

Dataset	Model	Scratch	Oracle
CIFAR-100	ResNet20	68.68	**90.60**
	ResNet32	70.90	**90.59**
	ResNet56	72.46	**90.99**
	ResNet110	74.11	**89.62**

Dataset	Model	Scratch	Oracle
ImageNet	ResNet18	69.76	**96.61**
	ResNet34	73.29	**96.08**
	ResNet50	76.14	**96.03**
	ResNet101	77.38	**95.82**

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 $\mathcal{P}_\mathcal{T}$ from the last layer of our teacher model, i.e., a probability distribution over different categories with a softmax layer: $\mathcal{P}_\mathcal{T} = \mathtt{Softmax}(\mathbf{o}_\mathcal{T}/\tau)$, where $\mathbf{o}_\mathcal{T}$ is the final output feature from the teacher model. We also use the temperature τ 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:

$$\mathcal{L}_{\text{distill}} = D_{KL}(\mathcal{P}_\mathcal{T} || \mathcal{P}_\mathcal{S}) \tag{3}$$

where $\mathcal{P}_\mathcal{S} = \mathtt{Softmax}(\mathbf{o}_\mathcal{S}/\tau)$ 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 $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{distill}} \tag{4}$$

where $\mathcal{L}_{\text{cls}} = H(\mathcal{P}_\mathcal{S}, y)$ is the cross-entropy loss for the classification task from the student model. We use hyperparameter α 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.

Model	# of Param	Scratch model	Oracle model
		ECE (↓)	ECE (↓)
ResNet18	11.2M	0.0327	**0.0311**
ResNet34	21.3M	0.0313	**0.0247**
ResNet50	23.7M	0.0284	**0.0216**
ResNet101	42.7M	0.0237	**0.0164**

model [7, 43, 50]. 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:

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

where $conf(b) = \frac{1}{|b|}\sum_{j \in b} p_j$ and $acc(b) = \frac{1}{|b|}\sum_{j \in b} \mathbb{1}(p_j = y_j)$ 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 [55]. It is generally observed that semantically similar inputs tend to evoke similar activation patterns in a trained model [18, 53]. 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-

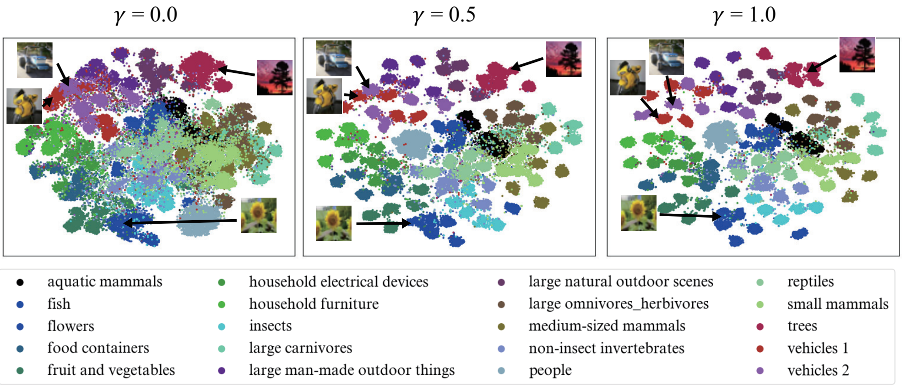

Figure 3: Visualizations by t-SNE [55] for output representation of the our Oracle teacher model (ResNet20) on CIFAR-100 dataset with varying γ in Eq. 2. For better understanding, we provide sample images and denote color coding points according to their superclass.

houette 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, $D_{ic}(k) = \frac{1}{|S_K|-1}\sum_{j\in S_K, k\neq j} d(k,j)$, and $D_{nc}(k)$ indicates *the nearest-cluster distance*, the smallest mean distance of k to all data in any other cluster, $D_{nc} = \underset{L\neq K}{min}\left(\frac{1}{S_L}\sum_{j\in S_L} d(k,j)\right)$.

$$Sil(k) = \frac{D_{nc}(k) - D_{ic}(k)}{max(D_{nc}(k), D_{ic}(k))} \tag{6}$$

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 γ in Eq. 2. As can be seen, silhouette scores for both super-class and sub-class increase as γ 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 γ is set too high, we observe a drop in the silhouette score. Therefore, we confirm that the optimal γ 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

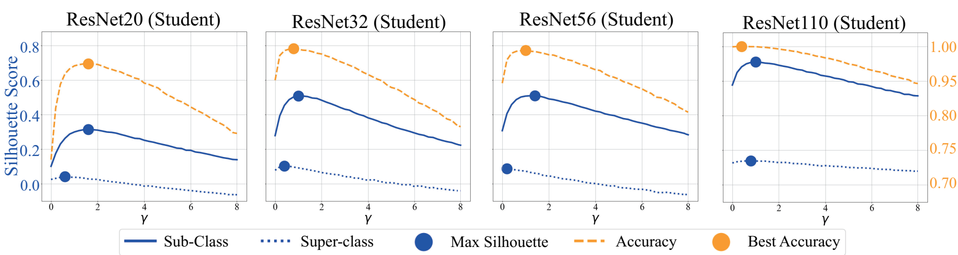

Figure 4: Variation of silhouette score and top-1 accuracy with varying γ in Eq. 2. The best and maximum scores are indicated by a circle.

Table 3: Comparison of entropy between confidence score about original data ($\mathbf{x}$) and Refined data ($\mathbf{x}^*$) on two datasets: CIFAR-10 and CIFAR-100.

Dataset	CIFAR-10				CIFAR-100			
Model	ResNet20	ResNet32	ResNet56	ResNet11 0	ResNet20	ResNet32	ResNet56	ResNet110
Original data ($\mathbf{x}$)	5.39	5.30	5.02	4.82	0.66	0.50	0.41	0.40
Reinforced data ($\mathbf{x}^*$)	5.01	4.87	4.60	4.25	0.37	0.27	0.22	0.21

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.

Dataset	CIFAR-10 dataset				CIFAR-100 dataset			
Teacher model = Student model	ResNet20 (90.31)	ResNet32 (90.88)	ResNet56 (91.62)	ResNet110 (92.06)	ResNet20 (69.26)	ResNet32 (71.14)	ResNet56 (72.34)	ResNet110 (74.31)
DKD [60]	90.27	91.07	92.25	91.99	69.42	72.87	74.94	76.21
ATT [59]	91.11	91.88	92.59	92.43	69.50	71.86	73.88	75.38
FitNet [42]	92.17	93.06	93.81	93.87	70.76	73.67	75.32	75.64
KD [18]	92.57	93.58	93.79	93.95	70.94	72.72	74.64	75.48
SAD [21]	92.71	93.61	93.52	93.37	70.52	73.34	75.01	76.36
Oracle (Ours)	**92.94**	**93.77**	**94.01**	**94.69**	**71.97**	**74.32**	**75.52**	**77.22**

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

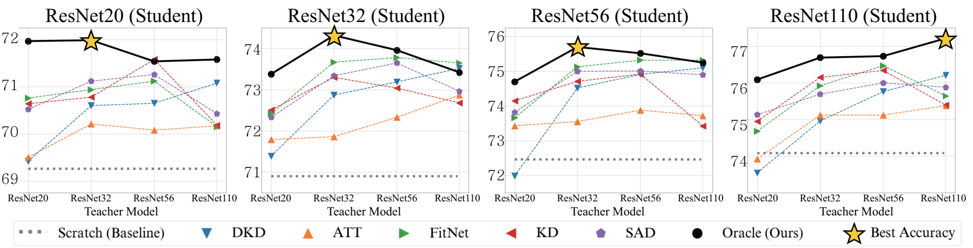

Figure 5: Comparison of accuracy of our Oracle and SOTA knowledge distillation methods on CIFAR-100 dataset. X-axis represents a different teacher model and Y-axis represents the accuracy. We denote the best performing method for each student model with a star symbol.

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.

Dataset	Tiny-ImageNet				ImageNet	
Teacher model	ResNet18 (59.83)		ResNet34 (61.50)		ResNet34 (73.31)	
Student model	ResNet18 (59.83)	ResNet34 (61.50)	ResNet18 (59.83)	ResNet34 (61.50)	ResNet18 (69.75)	ResNet34 (73.31)
DKD [60]	64.15	66.89	64.61	65.79	<u>71.85</u>	<u>74.84</u>
ATT [59]	63.58	65.04	64.59	65.08	71.33	73.83
FitNet [42]	65.56	<u>67.63</u>	65.16	67.28	71.11	74.72
KD [18]	65.13	67.43	<u>65.19</u>	66.22	71.34	74.69
SAD [21]	<u>65.78</u>	67.21	**66.33**	<u>67.79</u>	71.53	74.43
Oracle (Ours)	**66.25**	**67.99**	**66.33**	**67.87**	**72.02**	**74.85**

model. We further validate our findings by benchmarking performance on large-scale image classification datasets, including Tiny-ImageNet and ImageNet, to enhance the reliability of our experimental results. In Table 5, we observe ours outperforms other KD methods on Tiny-ImageNet dataset and ImageNet dataset.

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

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[2] Sarah Adel Bargal, Andrea Zunino, Vitali Petsiuk, Jianming Zhang, Kate Saenko, Vittorio Murino, and Stan Sclaroff. Guided zoom: Questioning network evidence for fine-grained classification. *arXiv preprint arXiv:1812.02626*, 2018.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[6] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[7] Hisham Daoud and Magdy Bayoumi. Deep learning approach for epileptic focus localization. *IEEE transactions on biomedical circuits and systems*, 14(2):209–220, 2019.

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

[9] Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. Learning credible deep neural networks with rationale regularization. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 150–159, 2019. doi: 10.1109/ICDM.2019.00025.

[10] Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients, 2020.

[11] Farzan Farnia, Jesse M Zhang, and David Tse. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

[12] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10705–10714, 2019.

[13] Mengya Gao, Yujun Shen, Quanquan Li, and Chen Change Loy. Residual knowledge distillation. *arXiv preprint arXiv:2002.09168*, 2020.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[16] Jin ha Lee, Ik hee Shin, Sang gu Jeong, Seung-Ik Lee, Muhamamad Zaigham Zaheer, and Beom-Su Seo. Improvement in deep networks for optimization using explainable artificial intelligence. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 525–530. IEEE, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[19] Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34:26726–26739, 2021.

[20] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *International Conference on Machine Learning*, pages 3030–3039. PMLR, 2019.

[21] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7945–7952, 2021.

[22] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[23] Minsoo Kang, Jonghwan Mun, and Bohyung Han. Towards oracle knowledge distillation with neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4404–4411, 2020.

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

[25] Chenxin Li, Mingbao Lin, Zhiyuan Ding, Nie Lin, Yihong Zhuang, Yue Huang, Xinghao Ding, and Liujuan Cao. Knowledge condensation distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 19–35. Springer, 2022.

[26] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017.

[27] Xuewei Li, Songyuan Li, Bourahla Omar, Fei Wu, and Xi Li. Reskd: Residual-guided knowledge distillation. *IEEE Transactions on Image Processing*, 30:4735–4746, 2021.

[28] Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1631. URL https://aclanthology.org/P19-1631.

[29] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[30] Dwarikanath Mahapatra, Alexander Poellinger, Ling Shao, and Mauricio Reyes. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2548–2562, 2021. doi: 10.1109/TMI.2021.3061724.

[31] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. Conditional teacher-student learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6445–6449. IEEE, 2019.

[32] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[33] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge into deep neural network via attention map, 2019.

[34] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[35] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

[36] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6553–6562, 2022.

[37] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

[38] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[39] Vineel Nagisetty, Laura Graves, Joseph Scott, and Vijay Ganesh. xai-gan: Enhancing generative adversarial networks via explainable ai systems, 2022.

[40] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13657–13665, 2021.

[41] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020.

[42] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[43] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[44] Bharat Bhusan Sau and Vineeth N Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016.

[45] Dominik Schiller, Tobias Huber, Florian Lingenfelser, Michael Dietz, Andreas Seiderer, and Elisabeth André. Relevance-Based Feature Masking: Improving Neural Network Based Whale Classification Through Explainable Artificial Intelligence. In *Proc. Interspeech 2019*, pages 2423–2427, 2019. doi: 10.21437/Interspeech.2019-2707.

[46] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[48] Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2591–2600, 2019. doi: 10.1109/ICCV.2019.00268.

[49] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[50] Meshal Shutaywi and Nezamoddin N Kachouie. Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, 23(6):759, 2021.

[51] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explanation-guided training for cross-domain few-shot classification, 2020.

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[53] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[54] Matt Turek. Explainable artificial intelligence (xai), 2021. URL https://www.darpa.mil/program/explainable-artificial-intelligence.

[55] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[56] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 478–495. Springer, 2022.

[57] Huan Wang, Suhas Lohit, Michael Jones, and Yun Fu. Knowledge distillation thrives on data augmentation. *arXiv preprint arXiv:2012.02909*, 2020.

[58] Leander Weber. Towards a more refined training process for neural networks: Applying layer-wise relevance propagation to understand and improve classification performance on imbalanced datasets. *Technische Universität Berlin*, 2020.

[59] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[60] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022.

[61] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[62] Andrea Zunino, Sarah Adel Bargal, Pietro Morerio, Jianming Zhang, Stan Sclaroff, and Vittorio Murino. Excitation dropout: Encouraging plasticity in deep neural networks, 2021.

[63] Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. Explainable deep classification models for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3233–3242, 2021.