iCD: An Implicit Clustering Distillation Method for Structural Information Mining (Student Abstract)

Xiang Xue[1], Yatu Ji[1,*], Qing-Dao-Er-Ji Ren[1], Bao Shi[1], Min Lu[1], Nier Wu[1], Xufei Zhuang[1], Haiteng Xu[1], Gan-Qi-Qi-Ge Cha[2]

[1]Inner Mongolia University of Technology, China
[2]Inner Mongolia Autonomous Region Water Conservancy Development Center, China
{20231100134, MLjyt, renqingln, kshibao, cslumin, wunier04, zxf, 20241100135}@imut.edu.cn, 1733860628@qq.com

Abstract

Logit Knowledge Distillation has gained substantial research interest in recent years due to its simplicity and lack of requirement for intermediate feature alignment; however, it suffers from limited interpretability in its decision-making process. To address this, we propose **i**mplicit **C**lustering **D**istillation (iCD): a simple and effective method that mines and transfers interpretable structural knowledge from logits, without requiring ground-truth labels or feature-space alignment. iCD leverages Gram matrices over decoupled local logit representations to enable student models to learn latent semantic structural patterns. Extensive experiments on benchmark datasets demonstrate the effectiveness of iCD across diverse teacher-student architectures, with particularly strong performance in fine-grained classification tasks—achieving a peak improvement of +5.08% over the baseline.

Code — https://github.com/maomaochongaa/iCD

Introduction

Knowledge distillation has become a crucial technique in model compression by transferring implicit knowledge embedded in pre-trained teacher models to enhance student networks' generalization performance (Gou et al. 2021; Gao et al. 2025). Among these, logit-level knowledge transfer has gained widespread attention in heterogeneous model compression tasks due to its low computational overhead and capacity for cross-architecture learning without requiring intermediate structural alignment.

Although existing logit distillation methods have achieved certain effectiveness in knowledge transfer, they solely rely on the global logit knowledge of inputs and struggle to fully transmit fine-grained semantic information, which may lead to suboptimal performance. To address this, SDD (Wei, Luo, and Luo 2024) attempts to refine knowledge through scale-level logit decoupling, but remains confined to the "learning-to-predict" paradigm in the output space and fails to model semantic structures in the representation space. To target this limitation, this paper proposes the iCD method, which aims to leverage scale-level decoupled logit outputs to further learn and transfer the teacher model's semantic structure-level knowledge, thereby enhancing the student model's discriminative capability and overall performance.

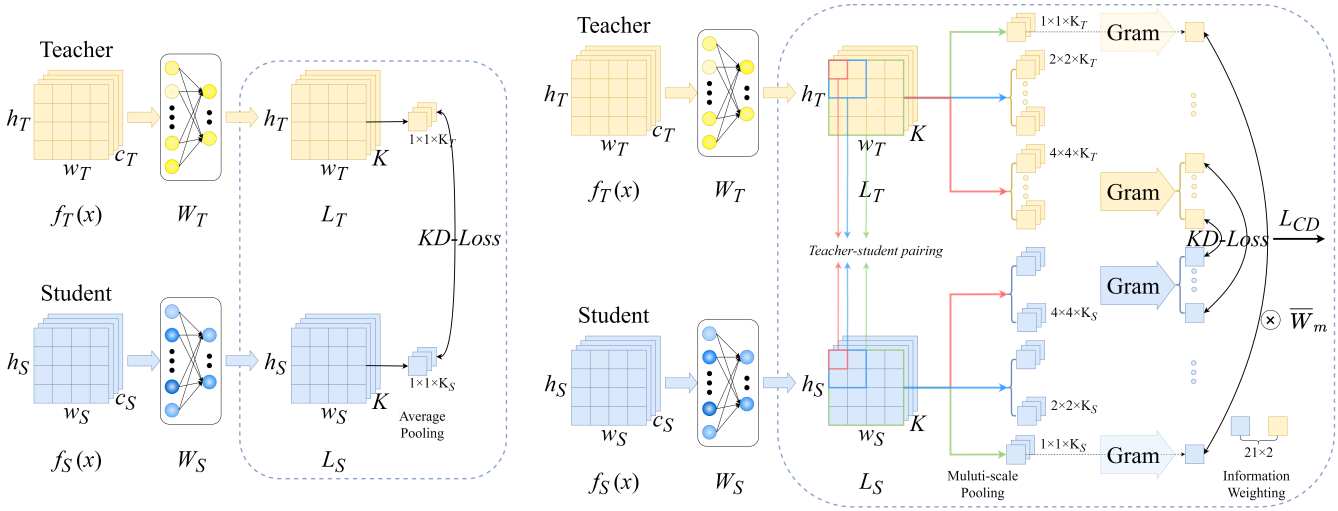
*Corresponding author.

Methodology

Notation. Given an input image x, let T and S denote the teacher network and student network, respectively. We divide these networks into two components: (i) A convolutional feature extractor f_{Net} where $Net = \{T, S\}$. The feature map from the penultimate layer is represented as $f_{Net}(x) \in R^{c_{Net} \times h_{Net} \times w_{Net}}$, where c_{Net} is the number of feature channels, and $h_{Net} \times w_{Net}$ denotes the spatial dimensions. (ii) A projection matrix $W_{Net}(x) \in R^{c_{Net} \times K}$ that maps the feature vectors extracted from $f_{Net}(x)$ to logits z^l_{Net} $(l = 1, 2, \ldots, K)$ corresponding to K categories. Let $f_{Net}(i, j) = f_{Net}(x)(:\, j, k) \in R^{c_{Net} \times 1 \times 1}$ denote the feature vector at spatial position (j, k) in $f_{Net}(x)$. According to the receptive field theory, $f_{Net}(j, k)$ corresponds to the representation of the region $(t_x, t_y, t_x + d, t_y + d)$ in x, where $t_x = d \cdot j$, $t_y = d \cdot k$, and d is the downsampling factor between the input and the final feature map. Define all scales m in the set $M = \{1, 2, 4, \ldots, w\}$, where each scale m partitions the feature map into $N_m = m^2$ non-overlapping cells. Let $C(m, n)$ denote the n-th spatial cell at scale m, where $n \in \{1, 2, \ldots, N_m\}$.

Distillation. As shown in Figure 1, the proposed iCD consists of two key components: *structured clustering* and *information weighting*. Specifically, given feature maps $f_T(x)$ and $f_S(x)$ extracted from the penultimate layers of the teacher and student models, respectively, our method forms logit pairs over corresponding spatial regions $C(m, n)$ and computes the similarity of their internal fine-grained representations. This helps students learn fine-grained logits with structured spatial distributions. Subsequently, we establish distillation pipelines for logits at each scale $m \in M$. Finally, information weighting reassigns weights across different scales, as the logits at larger m contain richer fine-grained information whose proportion increases accordingly. This guides the student network to focus precisely on local discriminative features.

In our method, $Z(m, n)$ represents the corresponding input region, and $\pi(m, n) \in R^{K \times 1 \times 1}$ denotes its logit output. We construct the structural matrix $G_{(m,n)}$ by computing the



(a) Conventional Knowledge Distillation

(b) Implicit Clustering Distillation

Figure 1: Visualization of Knowledge Distillation and Implicit Clustering Distillation.

Gram matrix of $\pi(m,n)^T$ and $\pi(m,n)^S$ across all scales, which captures the internal relational structure among logits. For each pair $G(m,n)^T$ and $G(m,n)^S$ at the same scale m, we compute the KL divergence loss between student and teacher models. Since logit outputs across scales vary in fine-grained knowledge content, applying uniform weights is suboptimal. We assign higher weights to finer-grained scales, encouraging the student to learn more discriminative local representations.

Experiments

Datasets and Implementation Details. Our experiments utilize the CIFAR-100 and CUB-200-2011 datasets. CIFAR-100 is adopted for evaluating standard image classification tasks, while CUB-200 is used for fine-grained image classification evaluation, which contains 200 bird species across distinct categories. For implementation details, we mostly follow the SDD (Wei, Luo, and Luo 2024) configuration.

Teacher	ResNet32x4	ResNet32x4	VGG13
Acc	66.17	66.17	70.19
Student	MobileNetV2	ShuffleNetV1	VGG8
Acc	40.23	37.28	46.32
SP	48.49	61.83	54.78
CRD	57.45	62.28	66.10
LDRLD	60.99	65.19	68.27
KD	56.09	61.68	64.18
SD-KD	60.51	65.46	67.32
iCD-KD	**61.17**	**65.88**	**68.52**
DKD	59.94	64.51	67.20
SD-DKD	62.97	65.58	**68.67**
iCD-DKD	**63.51**	**65.95**	68.31
NKD	59.81	64.00	67.16
SD-NKD	62.69	65.50	68.37
iCD-NKD	**63.55**	**67.10**	**68.95**

Table 1: Performance on the CUB-200 Dataset.

Results. As shown in Table 1, the proposed iCD method achieves strong performance across diverse teacher-student model combinations. Particularly in the most challenging cross-architecture distillation scenarios, the performance gains are notably pronounced: when transferring from ResNet32x4 to MobileNetV2, iCD-KD achieves a +5.08% improvement over KD. Notably, iCD-NKD outperforms the teacher model in the transfer from ResNet32x4 to ShuffleNetV1. These results conclusively validate that the proposed iCD method effectively mines and transfers refined structural knowledge from the teacher, significantly enhancing the capacity of the student model to assimilate complex, architecture-agnostic knowledge representations.

Conclusion

This paper proposes iCD, a novel knowledge distillation method designed to provide student models with an interpretable form of structured guidance. While SDD delivers decoupled logit supervision, iCD further mines the latent semantic structures within teacher and student logits. By encouraging student models to implicitly mimic the teacher's semantic structures (without relying on explicit class labels), iCD achieves more structure-aware knowledge transfer.

Acknowledgments

This study is supported by the Inner Mongolia Natural Science Foundation (2024MS06009), National Natural Science Foundation of China (62206138), Science and Technology Plan Project of Inner Mongolia Autonomous Region (2025YFHH0083, 2025YFHH0115), College Student Innovation and Entrepreneurship Training Program project (2024023003).

References

Gao, Z.; Han, S.; Zhang, X.; Xu, K.; Zhou, D.; Mao, X.; Dou, Y.; and Wang, H. 2025. Maintaining Fairness in Logit-based Knowledge Distillation for Class-Incremental Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16763–16771.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Wei, S.; Luo, C.; and Luo, Y. 2024. Scale Decoupled Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15975–15983.