Dean&Francis
ISSN 2959-6157

The Evolution of Knowledge Distillation in Image Classification Tasks

Tiejun Yin

School of Artificial Intelligence, Nanjing Normal University of Special Education
Nanjing, Jiangsu, China
073arfstu@stu.njts.edu.cn

Abstract:

In today's digital age, image classification plays a crucial role as a key task in the field of computer vision. Image classification tasks aim to accurately assign images to predefined categories, but training efficient models for high-precision classification operations using large-scale datasets remains challenging. To this end, researchers have discovered knowledge distillation strategies to achieve model performance compression. Knowledge distillation aims to transform complex models into lightweight ones through parameter optimization, enabling lightweight models to learn the capabilities of complex models from limited data without altering the original model structure, achieving excellent scalability. Knowledge distillation primarily involves the transfer of knowledge through three forms: model outputs, feature map matching, and structural knowledge. This paper primarily analyzes and discusses three aspects: decoupled knowledge distillation and decision boundary interpretation structures with an output-oriented focus, knowledge distillation based on classical feature-based attention transfer and Wasserstein distance, and relationship-based virtual distillation techniques and knowledge distillation that preserves Lipschitz continuity.

Keywords: Knowledge distillation, Computer vision, Model compression.

1. Introduction

The technological advancements and development of deep neural networks in the field of artificial intelligence have ushered in a historic transformation in computer vision. Deep learning acts as the driver in image classification. Compared with other models, convolutional neural network (CNN) uses convolutional kernels for local perception, which helps to reduce the amount of parameters and make the model more efficient. They extract features through multiple layers of convolution and pooling.

Application: Image classification is applied in the medical field, which includes medical imaging diagnostics based on X-ray images, CT images and MRI images. In the security field, such as facial recognition, dynamic monitoring, behavior scanning and analysis, etc. However, image classification tasks are constrained by data resources and model quality, and dataset imbalance also impacts the model's general-

ISSN 2959-6157

ization capability. Deep neural networks are often regarded as black boxes, limiting model reliability and rendering deep learning computations too complex to meet real-time requirements. The introduction of knowledge distillation effectively addresses the challenge of deploying overly large datasets required for training large models. By transferring knowledge from complex, massive yet high-performance teacher models to student models, it reduces the performance gap between them.

This paper employs the methodology proposed by classical knowledge distillation to progressively deepen the analysis to output-based structural, feature-level, and relational theoretical dimensions, providing a comprehensive overview of the primary strategies and key frameworks within knowledge distillation for the task of image classification.

2. Output-Oriented Knowledge Distillation

Knowledge Distillation (KD) is an outcome-oriented model compression technique originally proposed by Hinton et al. for image classification. It improves the performance of a small model (student model) by training it with the help of a large model (teacher model) [1]. This knowledge initially refers to the category probabilities obtained from the softmax layer of the teacher model. The student model takes the category probability vector as its learning target and mimics the teacher model's classification behavior. Distillation refers to raising the parameter temperature in the softmax layer to soften the output probability vector, making it easier for the student model to learn and thereby improving the generalization performance of the model. By improving the output-based knowledge distillation method and dividing the knowledge distillation into two dimensions, network performance and network compression, the method has improved logical coherence and interpretability [2].

2.1 Structural Innovation

As classical knowledge distillation is knowledge transferring based on the category probabilities output by the softmax layer, Borui Zhao et al. revisited the classical KD. Upon classical KD, they classified this classical KD into two approaches: Target-Class Knowledge Distillation (TCKD) aims at answering the target-class macro-level judgment question, i.e., whether the category is the target class[1]; and Non-Target-Class Knowledge Distillation (NCKD) focuses on how to suppress the non-target classes[3]. Specifically, the TCKD transmits prediction information for target categories, providing sample difficulty signals in the form of a binary classification task, while the NCKD reflects the similarity structure among negative categories. Due to coupling issues in the classic KD loss, which suppress NCKD effectiveness and prevent independent adjustment of TCKD and NCKD weights, decoupled knowledge distillation (DKD) is proposed to reduce coupling between the two and enhance efficiency. Based on classical knowledge distillation, DKD redefines the loss function as shown in Equation (1). Since the weights of NCKD are influenced by the target category probabilities of the teacher model, hyperparameters αand β are introduced to control the TCKD and NCKD components respectively, enabling truly independent adjustment of each part.

$$DKD = \alpha TCKD + \beta NCKD \tag{1}$$

2.2 The Essence of Distillation

Although researchers have explored various knowledge distillation techniques and their applications, such as Son, W et al. proposing a densely guided knowledge distillation technique that introduces multiple intermediate-scale teacher assistants for multi-stage knowledge distillation, enabling knowledge transfer between large-scale teacher models and small-capacity student models[4]. However, the understanding of the essence of knowledge distillation remains incomplete.To address this, Utkarsh Ojha et al. proposed a geometric interpretation framework based on decision boundaries. For the first time, they analyzed that latent knowledge is not only the inter-class relationships provided by soft labels but also a manifestation of the teacher's decision-making system. This confirmed that knowledge distillation essentially involves the student model learning the decision-making mechanism of the teacher model, thereby holistically shaping the student's decision-making behavior. It has been demonstrated that knowledge distillation also conveys implicit properties such as robustness, data invariance, and color constancy [5].

3. Feature-Based Knowledge Distillation

The classic KD method focuses solely on the final result, neglecting the characteristics of the teacher network's intermediate layers. This makes it increasingly difficult to optimize deeper networks as shallower student models mimic the teacher model's structure. To this end, Romero et al. utilized feature knowledge from intermediate layers to guide student network training. They selected the intermediate layer of the teacher network as the prompt layer and the intermediate layer of the student network as

the guidance layer. By adding a convolutional regressor on the guidance layer, they minimized the loss between the two networks. They adopted a strategy of training the student model with fixed teacher model parameters from the input to the guidance layer, followed by global network knowledge distillation training. They pioneered the introduction of prompt learning mechanisms to train deep models, breaking through the limitations of traditional output knowledge distillation [6]. Ziyao Guo et al. proposed a category-attention-transferring knowledge distillation (CAT-KD) with strong interpretability. They demonstrated that transferring only the category activation map (CAM) enhances the student model's ability to distinguish strong discriminative regions and guides it to focus on more important areas. The overall loss is the sum of cross-entropy loss and CAT loss with an introduced β balancing factor [7].

Although The model proposed by Romero et al. introduces a novel perspective by aligning intermediate-layer features between teacher and student networks to enhance student model performance, shallow feature alignment may lead to semantic inconsistencies when model structures differ significantly. As feature extraction grows more complex, the required transformation functions and alignment strategies also become more intricate, and processing higher-dimensional feature information demands greater computational resources.

Since the Leibler Divergence (KL-Div) in the same category has achieved certain applications in many fields and gained good performance, there are also certain limitations in KL-Div. KL-Div is only used for the same category comparison, can't be extended to inter-category comparison directly, and has certain problems when there is no overlapping region in the intermediate feature comparison. So Jiaming Lv et al. designed a knowledge distillation method based on Wasserstein Distance (WD) to compete with KL-Div, which can be divided into discrete Logits Distillation (WKD-L) and continuous Feature Distillation (WLD-F) [8].

The paper defines the WD as the minimum cost in . WKD-L employs a discrete WD to measure prediction discrepancies between models, quantifying category relationships through CKA to enable comparisons across different categories [9]. WKD-F models intermediate layer features using Gaussian distributions, matching feature maps via continuous WD while considering the Riemannian manifold geometry of positive definite symmetric matrices to effectively transfer deep feature knowledge. While WD-based knowledge distillation holds immense potential to overcome traditional method limitations, it still faces challenges such as high computational overhead and feature modeling constraints.

4. Relational Knowledge Distillation

Most feature-based and output-based knowledge distillation methods focus solely on knowledge within independent samples, whereas relational knowledge distillation places greater emphasis on structural knowledge within models and deep exploration of categorical relationships. Chuanguang Yang et al. discussed the general form of distillation losses as shown in Equation (2):

$$\mathcal{L}_{relation_{kd}}(F^S, F^T) = \sum_{ij} \mathcal{L}_{dis}(\psi^S(v_i^S, v_j^S), \psi^T(v_i^T, v_j^T)) \quad (2)$$

F^S, F^T representing the feature sets of the student model and teacher model respectively. v_i, v_j denote the feature embeddings of the i-th and j-th samples, respectively. ψ^S and ψ^T represent similarity measures for sample feature embeddings, $\mathcal{L}_{dis}$ while serves as the distance function for instance graph similarity [10].

Traditional relational knowledge distillation is weak in inducing relational matching, which leads to overfitting and interference from false information. Thus, its performance is far below instance-matching method. To conquer these problems, Weijia Zhan et al. proposed a new architecture, Virtual Relational Matching Knowledge Distillation (VRM) to help student model learn more informative affinity graphs which are rich in sample information, inter-class relations and inter-view structural relationships [11].

Firstly, they employ dense relationship graphs to learn inter-sample relationships from predicted logits. Then, they design a category-batch-level relationship graph structure to preserve response variations to learn structural knowledge. After that, they develop virtual graphs to learn virtual-real relationships. What is more, to alleviate false gradients, the affinity graph will be pruned twice to remove redundant edges at both source and target sides. At last, they combine huber loss with cross entropy loss.

A Novel Knowledge Distillation Approach learns virtual knowledge to improve model training revisits classical relational distillation and makes new improvements. For the first time, virtual relationship is introduced into graph structure of knowledge distillation. The exploration of relational distillation is reactivated and a larger search space is developed.

But the traditional feature based knowledge distillation only align the shallow-level knowledge, treating neural networks as black box and ignoring higher-level knowledge, i.e, functional features, which leads to the student directly imitating the teacher in a simple way. Thus, to bridge the gap, Shang Y. et al. proposed Lipschitz-Guided Knowledge Distillation (LONDON). They utilize Lip-

ISSN 2959-6157

schitz continuity as the knowledge to be transferred and obtain the knowledge transfer by minimizing the distance between Lipschitz constants of teacher-student networks [12].

Lipschitz continuity is typically defined as follows: for a real-valued function f:X→Y, where X and Y are metric spaces, if there exists a constant L ≥ 0 such that for all ∈ X, the difference between the two functions, i.e., $|f(x_1)-f(x_2)|$, and the rate of change between any two points is ≤ L, then L is called the Lipschitz constant. Although computing the Lipschitz constant is prohibitively difficult, the paper proposes approximating it using the transfer matrix of independent modules. If this matrix is normalized to become orthogonal, the spectral norm of the weight matrix can be obtained by calculating the maximum eigenvalue of the transfer matrix (avoiding direct computation of large-scale matrices). This approach approximates the Lipschitz constant for each module and employs a power iteration method for global network approximation.

The proposed LONDON knowledge distillation breakthrough the aforementioned limitations on solely considering shallow knowledge between features and outputs. Efficient approximation of constants through the power iteration method of transfer matrices, that serves as a solid theoretical support and available extension for relational distillation.

5. Comparative Analysis and Discussion

Table 1. Comparison of Pros and Cons of Knowledge Distillation.

	Output-Oriented Knowledge Distillation	Feature-Based Knowledge Distillation	Relational Knowledge Distillation
Advantages	Easy to implement and highly scalable, suitable for multiple tasks such as classification and detection.	Capable of capturing abstract details and conveying richer semantic information	Uncover deeper relationships, focus on model structural characteristics, and demonstrate strong generalization capabilities.
Disadvantages	Focusing solely on outcomes while neglecting feature information imposes significant limitations.	Cannot be applied to models with excessive structural differences; simple alignment operations introduce noise.	High implementation difficulty, significant computational resource consumption, unstable distillation results

Based on learning distillation and comparison in strengths and weaknesses (Table 1), knowledge network strategies to improve network effectiveness by exploring and innovating continuously have achieved remarkable performance as shown in Table 1.

Network knowledge ways from output distillation (a performance-centered approach which employs soft labels to express implicit knowledge), to feature distillation (which expresses abstract feature information for more concise performance compression), to relationship distillation (which establishes complicated relationships with models).

However, there are still many issues existing in knowledge distillation research, such as how to design effective metrics to evaluate knowledge distillation or not, how to implement tasks efficiently with huge and complicated data in ultra-high-precision scene, how to evaluate the transparency and regularity of knowledge distillation process, and how to solve the poor interpretability issue of some implicit information which may suppress distillation effectiveness. So, in the future, people should consider using more kinds of knowledge sources to improve the controllability of the model and satisfy the requirement of legitimacy and privacy. Unlike traditional transmission ways, students can learn how to communicate logically from teacher's network. Finally, people make a breakthrough in knowledge distillation in the view of green AI and sustainability.

6. Conclusions

This paper discusses the knowledge distillation methods for image classification in three different levels: output-based, feature-based and relationship-based. The simplest approach, which trains small models to mimic the outputs of large models, is straightforward and inexpensive but has inherent limitations.

Feature-based knowledge distillation uses feature knowledge for distillation, while only shallow-level feature information is aligned.

Unlike the previous studies based on single-sample, Relational knowledge distillation emboldens to propose distillation based on sample relationships. Theoretical refinement leads to a new attempt for knowledge transfer with

higher-level knowledge---function characteristics. Finally, the novel structured knowledge distillation effectively overcomes the weakness of classical knowledge distillation.

In contrast to classical knowledge distillation, in the future, the black-box models will be gradually tuned to be white-box models. Federated knowledge distillation, cross-modal joint distillation, multi-teacher knowledge fusion and PEFT technologies (including LoRA, Adapter etc.) will be widely used.

References

[1] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. arXiv Preprint arXiv:1503.02531, 2015.

[2] Si Zhaofeng, Qi Honggang. A Review of the Research and Application of Knowledge Distillation Methods. Journal of Image and Graphics, 2023, 28(09): 2817-2832.

[3] Zhao B, Cui Q, Song R, Qiu Y, Liang J. Decoupled Knowledge Distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11953-11962.

[4] Son W, Na J, Choi J, Hwang W. Densely Guided Knowledge Distillation Using Multiple Teacher Assistants. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 9395-9404.

[5] Ojha U, Li Y, Sundara Rajan A, Liang Y, Lee Y J. What Knowledge Gets Distilled in Knowledge Distillation? Advances in Neural Information Processing Systems, 2023, 36: 11037-11048.

[6] Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. FitNets: Hints for Thin Deep Nets. arXiv Preprint arXiv:1412.6550, 2014.

[7] Guo Z, Yan H, Li H, Lin X. Class Attention Transfer Based Knowledge Distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 11868-11877.

[8] Lv J, Yang H, Li P. Wasserstein Distance Rivals Kullback-Leibler Divergence for Knowledge Distillation. Advances in Neural Information Processing Systems, 2024, 37: 65445-65475.

[9] Cortes C, Mohri M, Rostamizadeh A. Algorithms for Learning Kernels Based on Centered Alignment. Journal of Machine Learning Research, 2012, 13(1): 795-828.

[10] Yang C, Yu X, An Z, Xu Y. Categories of Response-Based, Feature-Based, and Relation-Based Knowledge Distillation. In: Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems. Cham: Springer International Publishing, 2023: 1-32.

[11] Zhang W, Xie F, Cai W, Ma C. VRM: Knowledge Distillation via Virtual Relation Matching. arXiv Preprint arXiv:2502.20760, 2025.

[12] Shang Y, Duan B, Zong Z, Nie L, Yan Y. Lipschitz Continuity Guided Knowledge Distillation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10675-10684.