International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN(Online): 2347-5552, Volume-12, Issue-3, May 2024
https://doi.org/10.55524/ijircst.2024.12.3.17
Article ID IRP1455, Pages 106-112
www.ijircst.org

A Comprehensive Review of Knowledge Distillation- Methods, Applications, and Future Directions

Elly Yijun Zhu[1], Chao Zhao[2], Haoyu Yang[3], Jing Li[4], Yue Wu[5] and Rui Ding[6]

[1,6] San Francisco Bay University, USA
[2,3] Georgia Institute of Technology, USA
[4,5] Independent Researcher, USA

Correspondence should be addressed to Elly Yijun Zhu; ezhu9249@student.sfbu.edu

Received: 29 April 2024 Revised: 13 May 2024 Accepted: 26 May 2024

ABSTRACT- Knowledge distillation is a model compression technique that enhances the performance and efficiency of a smaller model (student model) by transferring knowledge from a larger model (teacher model). This technique utilizes the outputs of the teacher model, such as soft labels, intermediate features, or attention weights, as additional supervisory signals to guide the learning process of the student model. By doing so, knowledge distillation reduces computational resources and storage space requirements while maintaining or surpassing the accuracy of the teacher model. Research on knowledge distillation has evolved significantly since its inception in the 1980s, especially with the introduction of soft labels by Hinton and colleagues in 2015. Various advancements have been made, including methods to extract richer knowledge, knowledge sharing among models, integration with other compression techniques, and application in diverse domains like natural language processing and reinforcement learning. This article provides a comprehensive review of knowledge distillation, covering its concepts, methods, applications, challenges, and future directions.

KEYWORDS- Knowledge Distillation, Model Compression, Neural Networks, Soft Labels

I. INTRODUCTION

Knowledge distillation is a model compression technique aimed at transferring knowledge from a large model (referred to as the teacher model) to a smaller model (known as the student model), thereby enhancing the performance and efficiency of the student model. The fundamental idea behind knowledge distillation is to utilize the outputs of the teacher model (such as soft labels, intermediate features, or attention weights) as additional supervisory signals to guide the learning process of the student model. The advantages of knowledge distillation include reducing the computational resources and storage space requirements of the model while maintaining or even surpassing the accuracy of the teacher model, making it suitable for various machine learning tasks and domains.

Research on knowledge distillation began as early as the 1980s, but it wasn't until 2015 when Hinton and colleagues proposed the use of soft labels for knowledge distillation that it gained widespread attention [1]. This approach involves using the teacher model's probability vectors (adjusted by temperature) as soft labels, along with the true labels, as part of the student model's loss function. This way, the student model can learn not only the correct classes but also the confidence and uncertainty of the teacher model. This method has been proven effective in areas such as image classification and speech recognition.

With the development of deep learning, research on knowledge distillation has continued to advance, leading to many new methods and applications. For example, some methods attempt to extract richer knowledge from the teacher model's intermediate layers or other components (such as attention mechanisms [2] or convolutional kernels [3]), some methods consider knowledge sharing among multiple teacher or student models, some methods combine knowledge distillation with other model compression techniques (such as pruning [4] or quantization [5]) to further optimize the student model, and some methods apply knowledge distillation to fields like natural language processing, recommendation systems, and reinforcement learning. Additionally, theoretical analyses of knowledge distillation have also gained attention, aiming to reveal its essence and mechanisms to guide future research.

This article aims to provide a comprehensive review of the concepts, methods, and applications of knowledge distillation, while also discussing its challenges and future directions. The organization of this article is as follows: Section two introduces the basic framework and commonly used evaluation metrics of knowledge distillation; Section three reviews the main methods of knowledge distillation; Section four discusses the integration of knowledge distillation with other technologies; Section five introduces application scenarios of knowledge distillation; Section six discusses the challenges and future research directions of knowledge distillation; and finally, the conclusion.

II. KNOWLEDGE DISTILLATION

A. Basic Framework

Knowledge distillation is a model compression technique with the core idea of enabling a small student model to learn from a large teacher model's knowledge, thereby enhancing the student model's performance on the target task.

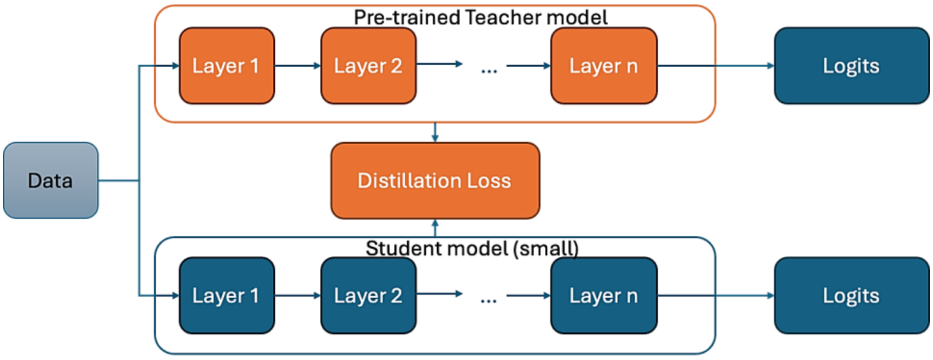

Figure 1: The basic framework of knowledge distillation

The basic framework of knowledge distillation is illustrated in Figure 1 and typically involves the following steps:

- Pretraining or Selecting a Suitable Teacher Model: The first step is to pretrain or select a teacher model suitable for the target task. The teacher model is usually a large deep neural network with high representational capacity and accuracy but comes with high computational and storage costs. Therefore, the teacher model may not be suitable for resource-constrained scenarios such as mobile devices or edge computing.
- Designing or Selecting a Smaller Student Model: The second step is to design or select a relatively smaller student model. The student model is typically a deep neural network with a structure similar to the teacher model but smaller in scale, possessing lower representational capacity and accuracy but also lower computational and storage costs. Therefore, the student model is more suitable for resource-constrained scenarios but requires knowledge distillation to enhance its performance.
- Conducting Knowledge Distillation with a Specific Dataset: The third step involves using a specific dataset (usually the training set or validation set of the target task) for knowledge distillation. The process of knowledge distillation can be seen as a form of supervised learning, where the teacher model serves as a soft label, the student model as a learner, and the dataset as input. The goal of knowledge distillation is to make the student model approximate the output distribution of the teacher model as closely as possible, thereby acquiring the teacher model's knowledge.
- Evaluating the Student Model's Performance: The fourth step is to evaluate the student model's performance on the target task using appropriate evaluation metrics to compare the differences between the student model and the teacher model, analyzing the effectiveness and influencing factors of knowledge distillation.

B. *Mechanism*

The mechanism of knowledge distillation involves introducing soft targets (such as the output distribution, intermediate feature representations, relational information, or structural information of the teacher model) to assist in training the student model, thereby transferring and refining the teacher model's knowledge. Soft targets can capture the teacher model's latent knowledge, such as predicted uncertainties, similarities, and correlations, which are beneficial for improving the student model's performance [6,7]. The process of knowledge distillation typically involves two loss functions: one based on hard targets for classification loss to ensure the correctness of the student model, and another based on soft targets for distillation loss to measure the difference between the student model and the teacher model. By balancing these two loss functions, the student model can maintain accuracy while approaching the teacher model's knowledge as closely as possible.

The following is the formula for the Knowledge Distillation loss function.

$$L = (1-\alpha)\cdot L_{\mathrm{CE}}(y,\widehat{y_s}) + \alpha\cdot T^2\cdot L_{\mathrm{KL}}\left(\sigma\left(\frac{\widehat{y_t}}{T}\right),\sigma\left(\frac{\widehat{y_s}}{T}\right)\right) \qquad (1)$$

L_{CE} is the cross-entropy loss function.

$$\mathrm{L_{CE}}(\mathrm{y},\widehat{y_s}) = -\sum_{\mathrm{i}} \mathrm{y_i}\log(\widehat{y_{\mathrm{s,\imath}}}) \qquad (2)$$

L_{KL} is Kullback-Leibler Divergence.

$$\mathrm{L_{KL}}(\mathrm{P},\mathrm{Q}) = \sum_{\mathrm{i}} \mathrm{P_i}\log\left(\frac{\mathrm{P_i}}{\mathrm{Q_i}}\right). \qquad (3)$$

σ is the softmax function, used to convert logits into a probability distribution.

$$\sigma(\mathrm{z_i}) = \frac{\mathrm{e^{z_i}}}{\sum_{\mathrm{j}}\mathrm{e^{z_j}}} \qquad (4)$$

The overall formula combination is as follows.

$$\mathrm{L} = (1-\alpha)\cdot\left(-\sum_{\mathrm{i}}\mathrm{y_i}\log(\widehat{y_{\mathrm{s,\imath}}})\right)$$
$$+\alpha\cdot\mathrm{T}^2\cdot\left(\sum_{\mathrm{i}}\sigma\left(\frac{\widehat{y_{\mathrm{t,\imath}}}}{\mathrm{T}}\right)\log\left(\frac{\sigma\left(\frac{\widehat{y_{\mathrm{t,\imath}}}}{\mathrm{T}}\right)}{\sigma\left(\frac{\widehat{y_{\mathrm{s,\imath}}}}{\mathrm{T}}\right)}\right)\right) \qquad (5)$$

C. *Evaluation Metrics*

To evaluate the effectiveness of knowledge distillation, various evaluation metrics are commonly used to compare the performance differences between the student model and the teacher model on the target task. The most common evaluation metric is accuracy, which measures the proportion of correctly predicted samples to the total samples. Accuracy reflects the model's generalization ability on the test set and is a primary evaluation metric for many machine learning tasks. However, accuracy alone may not fully reflect the purpose of knowledge distillation because it requires the student model not only to achieve or surpass the accuracy of the teacher model but also to inherit other features of the teacher model as much as possible, such as confidence, uncertainty, and robustness. Therefore, several other evaluation metrics have been proposed to assess the effectiveness of knowledge distillation from different perspectives. Here are some common evaluation metrics:

- Relative Error: Defined as the ratio of the error rate of the teacher model to the error rate of the student model, where E_s and E_t are the error rates of the student model and the teacher model, respectively.
  $$RelativeError = \frac{E_s}{E_t} \tag{6}$$
  Relative error measures how much the student model improves relative to the error rate of the teacher model. A value close to 1 indicates closeness between the two, with values smaller than 1 indicating the superiority of the student model over the teacher model.
- Distillation Factor: Defined as the ratio of the parameter count of the teacher model to the parameter count of the student model, where P_t and P_s are the parameter counts of the teacher model and the student model, respectively.
  $$DistillationFactor = \frac{P_t}{P_s} \tag{7}$$
  The distillation factor measures the efficiency of knowledge distillation in model compression, with larger values indicating higher compression ratios and lighter student models.
- Accuracy-Relative Error Curve: Plotting the performance of different student models on the test set against the relative error, with relative error on the x-axis and accuracy on the y-axis. This curve reflects the trade-off relationship between accuracy and relative error for different student models. Generally, models closer to the upper-left corner indicate better performance.
- Accuracy-Distillation Factor Curve: Plotting the performance of different student models on the test set against the distillation factor, with the distillation factor on the x-axis and accuracy on the y-axis. This curve reflects the trade-off relationship between accuracy and the distillation factor for different student models. Generally, models closer to the upper-right corner indicate better performance.
- Kullback-Leibler Divergence (KL Divergence): Measures the difference between the classification probability distributions of the teacher model and the student model, where P_t and P_s are the classification probability distributions of the teacher model and the student model, respectively.
  $$KL = D_{KL}(P_t \ || P_s) \tag{8}$$
  KL divergence quantifies how much the student model deviates from the teacher model on soft labels, with smaller values indicating closeness between the two and better learning of the teacher model's confidence and uncertainty.
- Correlation Coefficient: Measures the correlation between the classification probability distributions of the teacher model and the student model, where P_t and P_s are the classification probability distributions of the teacher model and the student model, respectively.
  $$Correlationcoefficient = \rho(P_t, P_s) \tag{9}$$
  The correlation coefficient quantifies the consistency of the student model with the teacher model on soft labels.

D. *Knowledge Forms*

The knowledge forms of knowledge distillation refer to the different types of knowledge accumulated by the teacher model during the training process and how this knowledge is transferred to the student model. Based on the source and representation of knowledge, knowledge distillation's knowledge forms can be categorized into the following four types:

- Output Feature Knowledge: This is the most common form of knowledge and is also used in classical knowledge distillation methods. Output feature knowledge refers to the probability distribution vector generated by the teacher model at the output layer (usually the softmax layer), also known as soft labels. Soft labels contain more information compared to hard labels and can reflect the teacher model's confidence and preferences for different classes. By using soft labels as additional supervision signals, the student model can better fit the data distribution and improve generalization ability. [8-13]
- Intermediate Feature Knowledge: Apart from the output layer, the teacher model also produces valuable feature representations at intermediate layers, reflecting its abstract understanding and encoding of input data. Intermediate feature knowledge refers to using the feature representations from the intermediate layers of the teacher model to guide the training of the student model. Aligning the intermediate layers of the teacher model and the student model can help the student model converge faster and learn more effective feature representations. [14-18]
- Relationship Feature Knowledge: In addition to individual feature representations, the teacher model also contains implicit relationship information internally, such as similarity between samples or dependencies between classes. Relationship feature knowledge refers to using this internal relationship information from the teacher model as soft targets to assist the student model in learning. Aligning the relationship information between the teacher model and the student model can help the student model better capture the structure and semantic information of the data and enhance the model's robustness. [19-26]

Structural Feature Knowledge: Apart from content information, the structure of the teacher model also contains knowledge, such as graph structures, hierarchical structures, attention mechanisms, etc. Structural feature knowledge refers to the student model improving its performance by learning the structural information from the teacher model. By making the structure of the teacher model and the student model similar or compatible, the student model can better utilize the computational resources and optimization strategies of the teacher model, improving flexibility and scalability. [27-29]

III. KNOWLEDGE DISTILLATION METHODS

Knowledge fusion refers to integrating the knowledge of multiple teacher models into a single student model. This approach leverages the complementarity of different teacher models to enhance the coverage and diversity of the student model, while also reducing bias and noise from a single teacher model.

A. Multi-Teacher Learning

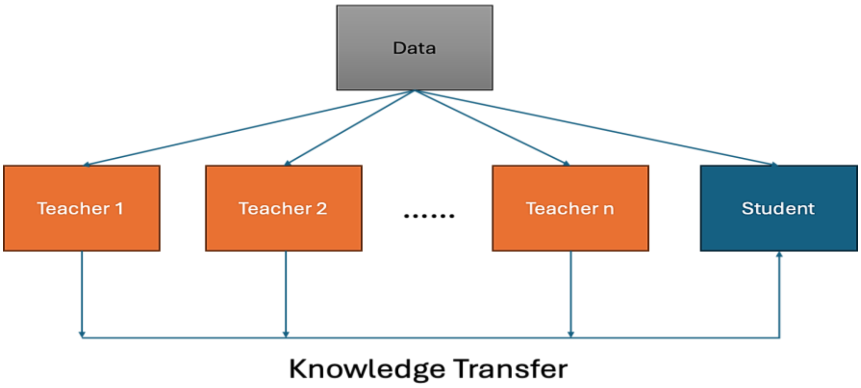

Figure 2: framework of Multi-teacher learning

Multi-teacher learning framework demonstrated in Figure 2 involves using the knowledge of multiple teacher models simultaneously to train a student model, enhancing robustness and generalization performance. This method allows the student model to learn knowledge from multiple perspectives, avoiding overfitting or underfitting, and can also increase the adaptability and transferability of the student model.

You et al. [30] proposed a strategy to integrate the outputs of multiple teacher models and used a weighted averaging method to generate the learning targets for the student model. This approach has demonstrated excellent performance across various tasks. Liu et al. [31] introduced an intermediate representation layer to facilitate more effective knowledge transfer from multiple teacher models to the student model. This method has shown its superiority on multiple benchmark datasets. Park et al. [20] utilized the relational information between teacher models to guide the training of the student model. This method not only focuses on the outputs of individual models but also considers the relationships between models, thereby further enhancing the performance of the student model.

B. Teacher Assistant

Teacher assistant involves introducing auxiliary tasks to enhance the knowledge of teacher models before transferring it to the student model. This method allows teacher models to learn more relevant knowledge during training, thereby improving the effectiveness and quality of the student model and reducing the complexity gap between teacher and student models.

Mirzadeh et al. [32] introduced one or more assistant models between the teacher and student models to transfer knowledge in stages. Each assistant model is responsible for learning from the previous model (either the teacher or the previous assistant) and passing the knowledge to the next model (either the next assistant or the student). In multiple benchmarks, this method significantly improved the performance of the student model, demonstrating the effectiveness of the teacher assistant model in knowledge distillation. Gou et al. [33] summarized the role and advantages of the teacher assistant model in knowledge distillation, discussing its contributions to improve knowledge transfer efficiency and enhancing student model performance. They also highlighted the challenges faced by the teacher assistant model in the field of knowledge distillation, such as the selection and optimization of assistant models, and proposed potential future research directions and improvement strategies.

C. Cross-Modal Distillation

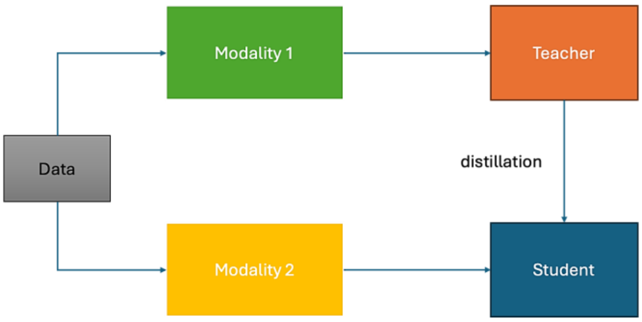

Figure 3: framework of Cross-modal distillation

Cross-modal distillation illustrated in Figure 3 involves knowledge distillation across different data modalities to improve the model's multimodal performance. In this framework, data is processed through distinct modal pathways – Modality 1 and Modality 2. Modality 1 informs the teaching strategy employed by the Teacher model, which in turn guides the learning of the Student model. Modality 2 directly influences the Student model. This approach enables the student model to leverage knowledge from teacher models across different modalities, enhancing the student model's expressive and reasoning abilities, and expanding its application scenarios and functionalities.

Yang [33] et al. investigated cross-modal distillation in text-to-image generation tasks and proposed a Dual-Modality Knowledge Distillation [34] method. By simultaneously optimizing text and image generation models, they improved the quality and consistency of image generation. Kim et al. introduced a knowledge distillation method from speech to text, named Speech2Text Distillation [35], leveraging pre-trained speech recognition models to enhance text generation models. They significantly improved the performance of speech-to-text tasks through cross-modal distillation.

D. Mutual Distillation

Mutual distillation involves the mutual transfer of knowledge between teacher and student models to collectively improve performance. This method allows teacher and student models to reference each other, improving their consistency and similarity, and promoting mutual learning and improvement. Zhang et al. [36] proposed that models are trained independently on the dataset to learn the features and patterns of the raw data. Then, the feature representations of the models are passed to other models, allowing them to learn each other's feature representations. Based on the exchanged feature information, each model updates its parameters to better capture the data's features and improve model performance. Chen et al. [37] proposed that knowledge exchange between peer models can be achieved by passing probability distributions, feature representations, or model parameters. Based on the knowledge from peer models, updates and optimizations are made to the target model to enhance its generalization ability and performance.

E. Lifelong Distillation

Lifelong distillation utilizes the knowledge of historical models to assist in the training of the current model, maintaining model performance stability. This method allows the current model to gain experience and insights from historical models, improving its rapid convergence and adaptation to new data, and preventing catastrophic forgetting or performance degradation issues.

Li et al. [38] proposed a knowledge distillation-based method, which transfers knowledge from a teacher model to a student model, thereby achieving the goal of retaining old task knowledge while learning new tasks. By designing appropriate loss functions and model update strategies, this method can effectively avoid conflicts between new and old knowledge, thus achieving the "learning without forgetting" effect. Hou et al. [39] introduced a progressive distillation method, gradually distilling old knowledge into a new model, thus achieving the goal of lifelong learning. What sets this article apart is the introduction of a retrospection mechanism, where during the learning of new tasks, old task knowledge is revisited and strengthened to improve the model's balance and stability between old and new tasks.

F. Self-Distillation

Self-distillation optimizes the model itself by learning its own soft targets to enhance generalization capability. This method allows the model to generate smoother and more flexible targets during training, thereby improving uncertainty and robustness, and reducing variance and noise. Yun et al. [40] proposed a method for regularizing class-wise predictions via self-knowledge distillation. Through self-distillation, the model can learn the correlation and distinctiveness between classes, thereby improving the classification performance across different categories. Xu et al. [41] introduced a self-distillation method guided by data distortion for deep neural networks. By incorporating data distortion, the model can better learn the features of data and enhance its performance through self-distillation. Nie et al. [42] presented a dynamic kernel distillation method for efficient pose estimation in videos. Through self-distillation, the model can learn the representation of dynamic features in videos, enhancing the efficiency of pose estimation.

IV. APPLICATION SCENARIOS

Knowledge distillation is widely used in fields such as image classification, object detection, semantic segmentation, natural language processing, etc., to reduce computational and storage costs while maintaining model accuracy, making it particularly suitable for mobile and edge computing scenarios. Here are some specific application scenarios:

A. Image Classification: By applying knowledge distillation, large-scale image classification models like ResNet [43], VGG [44], etc., can be compressed into smaller models like MobileNet [45], ShuffleNet [46], etc., enabling fast and accurate image classification on mobile devices.

B. Object Detection: Knowledge distillation can compress high-performance object detection models like Faster R-CNN [47], YOLO [48], etc., into lightweight models like SSD [49], Tiny YOLO [50], etc., allowing real-time and precise object detection in resource-constrained environments.

C. Semantic Segmentation: Complex semantic segmentation models like DeepLab [51], PSPNet [52], etc., can be compressed into simpler models like SegNet [53], ENet [54], etc., through knowledge distillation, achieving efficient and accurate semantic segmentation on edge devices.

D. Natural Language Processing: Large-scale natural language processing models like BERT [55], GPT [56], etc., can be compressed into smaller models like DistilBERT [57], MobileBERT [58], etc., through knowledge distillation, enabling high-performance and low-resource natural language processing across various tasks.

V. CHALLENGES AND FUTURE DIRECTIONS

As an advanced technology, knowledge distillation also faces some challenges and issues that require further research to explore and solve. These mainly include:

G. Model Robustness: Improving the student model's robustness to noise and interference is one of the future challenges. Since knowledge distillation is based on learning from soft targets, the student model may become overly sensitive to errors or uncertainties from the teacher model, affecting its robustness. Future research can explore methods to enhance the model's robustness during knowledge distillation, such as adversarial training, noise injection, confidence filtering, etc.

H. Cross-Domain Knowledge Transfer: Effectively transferring knowledge between different data distributions and domains is another research direction for the future. When the data distributions or domains of the teacher and student models are inconsistent, knowledge mismatch or adaptation issues may arise, affecting model performance. Future research can explore methods to enhance the model's cross-domain adaptation during knowledge distillation, such as domain adaptation, domain alignment, domain generation, etc.

I. Structured Knowledge Distillation: Leveraging internal structural information of models for knowledge distillation is another future research direction. Since knowledge distillation operates at the feature level, it may overlook structural information within the model, such as graph structures, hierarchical structures, etc., which can be beneficial for model performance and understanding. Future research can explore methods to incorporate structured knowledge into knowledge distillation, such as graph convolutional networks, attention mechanisms, structural similarity, etc.

J. Multimodal Fusion: Integrating visual, linguistic, and other multimodal information into knowledge distillation to enhance the model's multimodal performance is another research direction for the future. Since knowledge distillation primarily focuses on learning from a single modality, it may overlook the complementarity and enhancement between different modalities, which can be beneficial for model performance and generalization. Future research can explore methods to incorporate multimodal knowledge into knowledge distillation, such as cross-modal distillation, multimodal fusion, multimodal generation, etc.

VI. CONCLUSION

Knowledge distillation is an effective technique for model compression and optimization, transferring knowledge from teacher models to student models to improve performance. However, it also faces challenges and issues such as balancing goals and constraints, expanding scenarios and scopes, etc. Addressing these issues through future research will further promote the development and application of knowledge distillation.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

REFERENCES

[1] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503. 02531, 2015

[2] Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers//Proceedings of the Advances in Neural Information Processing Systems. 2020:1-15

[3] Wang Z, Deng Z, Wang S. Accelerating convolutional neural networks with dominant convolutional kernel and knowledge pre-regression//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 533-548

[4] Li T, Li J, Liu Z, Zhang C. Few sample knowledge distillation for efficient network compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 14639-14647

[5] Polino A, Pascanu R, Alistarh D. Model compression via distillation and quantization//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-21

[6] Tang Z, Wang D, Zhang Z. Recurrent neural network training with dark knowledge transfer//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, 2016: 5900-5904

[7] Yuan L, Tay F E H, Li G, Wang T, Feng J. Revisiting knowledge distillation via label smoothing regularization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3903-3911

[8] Chen G, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation//Proceedings of the 30th International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 742-751

[9] Wang T, Yuan L, Zhang X, Feng J. Distilling object detectors with fine-grained feature imitation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4933-4942

[10] Hou Y, Ma Z, Liu C, Hui T-W, Loy C C. Inter-Region affinity distillation for road marking segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12486-12495

[11] Liu Y, Chen K, Liu C, Qin Z, Luo Z, Wang J. Structured knowledge distillation for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2604-2613

[12] Takashima R, Sheng L, Kawai H. Investigation of sequence-level knowledge distillation methods for CTC acoustic models//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, 2019: 6156-6160

[13] Huang M, You Y, Chen Z, Qian Y, Yu K. Knowledge distillation for sequence model//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India, 2018: 3703-3707

[14] Gotmare A, Keskar N S, Xiong C, Socher R. A closer look at deep learning heuristics: learning rate restarts, warmup and distillation//Proceedings of the 7th International Conference on Learning Representations. New Orleans,USA, 2019:1-16

[15] Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. Fitnets: hints for thin deep nets//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015:1-13

[16] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017:1-13

[17] Li X, Xiong H, Wang H, Rao Y, Liu L, Huan J. Delta: deep learning transfer using feature map with attention for convolutional networks//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019:1-13

[18] Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 268- 284

[19] Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: fast optimization, network minimization and transfer learning// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4133-4141

[20] Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3967-3976

[21] Srinivas S, Fleuret F. Knowledge transfer with Jacobian Matching//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 4723-4731

[22] Lee S H, Kim D H, Song B C. Self-supervised knowledge distillation using singular value decomposition//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 339-354

[23] Chen Y, Wang N, Zhang Z. Darkrank: accelerating deep metric learning via cross sample similarities transfer//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2852-2859

[24] Peng B, Jin X, Liu J, Zhou S, Wu Y, Liu J, Zhang Z, Liu Y. Correlation congruence for knowledge distillation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5006-5015

[25] Lee S, Song B C. Graph-based knowledge distillation by multihead attention network//Proceedings of the 30th British Machine Vision Conference. Cardiff, UK, 2019: 141

[26] Bajestani M F, Yang Y. Tkd: Temporal knowledge distillation for active perception//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 953-962

[27] Liu Y, Shu C, Wang J, Shen C. Structured knowledge distillation for dense prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, (42): 1-15

[28] Xu X, Zou Q, Lin X, Huang Y, Tian Y. Integral knowledge distillation for multi-Person pose estimation. IEEE Signal Processing Letters, 2020, (27): 436-440

[29] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[30] You, S., Xu, C., Xu, C., & Tao, D. Learning from multiple teacher networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 1285-1294

[31] Liu, I. J., Peng, J., & Schwing, A. G. Knowledge flow: Improve upon your teachers. 2019: arXiv preprint arXiv:1904.05878.

[32] Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., & Ghasemzadeh, H. (2020, April). Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 04, pp. 5191-5198).

[33] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.

[34] Yang, G., Tang, Y., Wu, Z., Li, J., Xu, J., & Wan, X. (2024, April). DMKD: Improving Feature-Based Knowledge Distillation for Object Detection Via Dual Masking Augmentation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3330-3334). IEEE.

[35] Kim, S., Kim, G., Shin, S., & Lee, S. (2021, June). Two-stage textual knowledge distillation for end-to-end spoken language understanding. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7463-7467). IEEE.

[36] Zhang Y, Xiang T, Hospedales T M, Lu H. Deep mutual learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4320-4328

[37] Chen D, Mei J P, Wang C, Feng Y, Chen C. Online knowledge distillation with diverse peers//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3430- 3437

[38] Li Z, Hoiem D. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935-2947

[39] Hou S, Pan X, Change Loy C, Wang Z, Lin D. Lifelong learning via progressive distillation and retrospection//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 437-452

[40] Yun S, Park J, Lee K, Shin J. Regularizing class-wise predictions via self-knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13876-13885

[41] Xu T B, Liu C L. Data-distortion guided self-distillation for deep neural networks//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 5565-5572

[42] Nie X, Li Y, Luo L, Zhang N, Feng J. Dynamic kernel distillation for efficient pose estimation in videos//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 6942-6950

[43] Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

[44] Sengupta, A., Ye, Y., Wang, R., Liu, C., & Roy, K. (2019). Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in neuroscience, 13, 95.

[45] Sinha, D., & El-Sharkawy, M. (2019, October). Thin mobilenet: An enhanced mobilenet architecture. In 2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON) (pp. 0280-0285). IEEE.

[46] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848-6856).

[47] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[48] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. Procedia computer science, 199, 1066-1073.

[49] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 21-37). Springer International Publishing.

[50] Fang, W., Wang, L., & Ren, P. (2019). Tinier-YOLO: A real-time object detection method for constrained environments. Ieee Access, 8, 1935-1944.

[51] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[52] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[53] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[54] Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.

[55] Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

[56] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

[57] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[58] Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.