FLRKD: Relational Knowledge Distillation Based on Channel-wise Feature Quality Assessment

Zeyu An[a], Changjian Deng[a], Wanli Dang[a,b], Zhicheng Dong[c], Qian Luo[b], Jian Cheng[a]

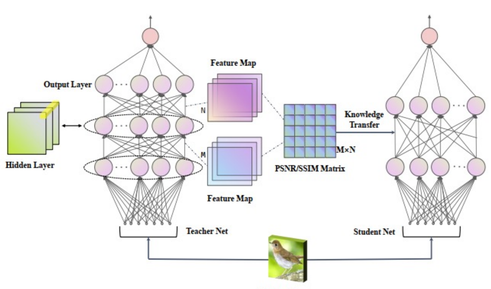
[a] School of Information and Communication Engineering, University of Electronic Science and Technology of China
[b] The Second Research Institute of Civil Aviation Administration of China
[c] School of Information Science and Technology, Tibet University

Abstract

With the increasing computational power of computing devices, the pre-training of large deep-learning models has become prevalent. However, deploying such models on edge devices with limited memory and computing power remains a significant challenge. To address this issue, this study proposes a novel knowledge distillation approach called Feature-level Relationship-based Knowledge Distillation (FLRKD). The proposed approach employs image quality similarity assessment to distill knowledge from a pre-trained model into smaller models that are suitable for deployment on edge devices. FLRKD utilizes peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) between feature maps of different hidden layers as relational knowledge to enhance the classification accuracy of student models. Moreover, the proposed approach includes an effective loss function that accelerates the convergence of the knowledge distillation algorithm. Additionally, a regressor is introduced to address the issue of inconsistent feature map spatial size between teacher and student models in heterogeneous scenarios. Comparative and ablation experiments demonstrate the superiority of FLRKD over mainstream knowledge distillation methods in terms of higher classification accuracy (up to 4%) and faster convergence rates. Notably, the proposed approach achieves significant improvement in classification accuracy (up to 3%) even in heterogeneous scenarios compared to existing state-of-the-art methods.

Highlights of this work

- We propose a new type of knowledge that is characterized by the PSNR matrix and SSIM matrix computed between different channels in different hidden layers, which has a better effect on guiding student model training than the knowledge used in the current mainstream relationship-based knowledge distillation methods.
- Employing the proposed refinement approach to obtain initial weights can effectively enhance the performance of compact neural networks, while also yielding a quicker convergence rate.
- Unlike other relationship-based knowledge distillation methods, our approach can also be applied when the teacher and student models are heterogeneous. Even if the structure of the student model differs from that of the teacher model, our proposed method can significantly improve the performance of the student model.



Methodology & Network Architecture

- **Feature-level Relationship-based Knowledge**: Response-based knowledge overlooks the knowledge contained in the hidden layer of the neural network. Feature-based knowledge distillation methods imitate the intermediate results of the feature layer of the teacher network but fail to consider the spatial relationship between shallow and deep layers. To address these issues, we propose a knowledge distillation method based on the flow of problem-solving processes represented by PSNR and SSIM matrices between feature maps from different layers. Our method avoids the problem of selecting effective feature maps and accurately represents the hierarchical relationships in the model.

$$P_{a/b,c/d}(x;W) = 10 \cdot \log_{10}\left(\frac{MAX}{\frac{1}{hw}\sum_{s=0}^{h-1}\sum_{t=0}^{w-1}\left\|F_{s,t,i}^{T/S}(x;W) - F_{s,t,j}^{T/S}(x;W)\right\|}\right)$$

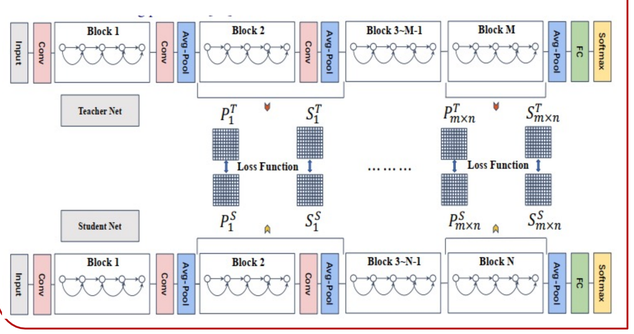
$$S_{a/b,c/d}(x;W) = \left[l\left(F_{s,t,i}^{T/S}(x;W), F_{s,t,j}^{T/S}(x;W)\right)\right]^{\alpha} \times \left[c\left(F_{s,t,i}^{T/S}(x;W), F_{s,t,j}^{T/S}(x;W)\right)\right]^{\beta} \times \left[s\left(F_{s,t,i}^{T/S}(x;W), F_{s,t,j}^{T/S}(x;W)\right)\right]^{\gamma}$$

- **Loss for the FLRKD**: The overall loss function contains an image spatial size correction regressor as well as PSNR and SSIM matrix correction losses.

$$\mathcal{L}_{PSNR}(W_T, W_S) = \frac{1}{N}\sum_{x}\sum_{i=1}^{\eta}\lambda_i \times \left\|P_i^T(x;W_T) - P_i^S(x;W_S)\right\|_2$$

$$\mathcal{L}_{SSIM}(W_T, W_S) = \frac{1}{N}\sum_{x}\sum_{i=1}^{\eta}\lambda_i \times \left\|S_i^T(x;W_T) - S_i^S(x;W_S)\right\|_2$$

Figure 2: Framework of our FLRKD algorithm when the ResNet structure is used in the teacher and student networks.



Results

Table 1 presents average accuracy results for knowledge distillation algorithms on the CIFAR-100[1] dataset using identical architectures for both teacher and student networks. Although FLRKD is not optimal for all knowledge distillation algorithms in a few cases, the starting point of this work is to propose a more efficient relationship-based knowledge distillation algorithm and to provide a feasible new idea for relationship-based knowledge distillation algorithms.

Teacher Student	wrn-40-2 wrn-16-2	wrn-40-2 wrn-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32×4 resnet8×4	vgg13 vgg8
Teacher Model	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student Model	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNets	73.55	72.31	69.21	69.00	71.10	73.49	71.07
SP	73.38	72.40	69.63	70.12	72.70	72.93	72.71
RKD	73.32	72.18	69.48	69.32	71.79	71.88	71.53
PKT	74.64	73.49	70.33	70.31	72.54	73.69	72.89
FT	73.12	71.60	69.76	70.26	72.37	72.81	70.58
FSP	72.78	72.51	69.91	70.07	71.83	72.58	70.28
QuEST	74.12	73.51	69.86	69.99	72.74	73.31	72.92
SimKD	74.76	73.57	70.25	70.63	73.06	**74.52**	73.14
TDD	74.73	73.44	70.00	70.52	72.98	73.10	72.94
DKD	**75.02**	73.89	**70.74**	70.81	73.15	74.48	**73.96**
FLRKD(ours)	74.14	73.62	70.35	70.97	73.22	73.56	73.17
FLRKD+KD	74.96	**74.06**	70.68	**71.15**	**73.24**	73.61	73.83
w/o SSIM	73.44	72.97	69.76	70.56	73.02	73.39	71.60
w/o PSNR	73.92	73.09	69.41	70.83	73.11	73.40	72.06

Table 1: Test accuracy of student networks on CIFAR-100[1] of several distillation methods (ours is FLRKD). We note that the FLRKD algorithm has the best performance among the relationship-based knowledge distillation algorithms, and we have marked relationship-based knowledge distillation algo- rithms in purple font in the table. The data in bold in the table correspond to the highest classification accuracy and the data in green font in the table correspond to the second-best classification accuracy obtained by using different knowledge distillation algorithms with the same teacher-student network frameworks and the same experimental conditions.

Pre-training deep neural networks has become a trend in recent years, but as network size increases, training time also increases[2]. Despite this, researchers continue to develop better models due to the excellent performance of deep neural networks in various fields. Therefore, there is a growing demand for fast and lightweight technology. In our proposed technique, we used one teacher network to generate several student networks, aiming to achieve similar performance with less training time than the normal procedure[2,3].

Algorithms	Original Model	Distillation Model	Distillation time
FSP	98.0 MB	10.9 MB	104 min
RKD	98.0 MB	9.9 MB	71 min
TDD	98.0 MB	13.0 MB	105 min
DKD	98.0 MB	9.3 MB	97 min
FLRKD(ours)	98.0 MB	**11.3 MB**	**82 min**

The experimental results in Table 3 show that the FLRKD algorithm achieves accuracy comparable to state-of-the-art knowledge distillation methods but with a shorter distillation time. In comparison to TDD[4], and DKD[5], FLRKD exhibits superior convergence speed and achieves desirable accuracy.

Table 2 presents the average accuracy of different knowledge distillation methods on CIFAR-100 dataset with varying teacher-student architectures. Based on the experimental findings presented in Table 2, it is observed that the FLRKD combined with the KD algorithm achieves generally superior accuracy than other knowledge distillation methods, even in scenarios where the teacher-student model has distinct architectures.

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32×4 ShuffleNetV1	resnet32×4 ShuffleNetV2	wrn-40-2 ShuffleNetV1
Teacher Model	74.64	79.34	79.34	79.42	79.42	75.61
Student Model	64.60	64.60	70.36	70.50	71.82	70.50
KD	67.37	67.35	73.81	74.07	74.45	74.83
FitNets	64.14	63.16	70.69	73.59	73.54	73.73
SP	66.30	68.08	73.34	73.48	74.56	74.52
RKD	64.52	64.43	71.50	72.28	73.21	72.21
PKT	67.13	66.52	73.01	74.10	74.69	73.89
FT	61.78	60.99	70.29	71.75	72.50	72.03
QuEST	67.92	67.76	73.80	74.58	74.65	75.45
SimKD	68.45	68.33	74.33	75.12	75.91	76.36
TDD	68.61	68.09	74.28	74.96	74.88	75.47
DKD	69.43	68.54	74.32	**75.33**	75.83	**76.59**
FLRKD(ours)	68.59	68.11	74.42	75.17	75.77	76.02
FLRKD+KD	**69.81**	**69.24**	**75.28**	75.32	**76.00**	**76.59**
w/o SSIM	68.01	67.52	73.60	74.48	74.75	75.14
w/o PSNR	68.13	67.99	74.03	75.03	75.18	75.74

Table 2: Test accuracy of student networks on CIFAR100 of a number of distillation methods (ours is FLRKD) for transfer across different teacher and student architectures. Importantly, some methods that require very similar student and teacher architectures perform quite poorly. E.g. FSP cannot even be applied. FLRKD can also be adapted to the Teacher-Student model using different architectures by introducing regressors. The meanings of different fonts in Table 2 are consistent with those in Table 1.

Conclusions & Future Work

In this work, we propose a new method of relationship-based knowledge distillation, based on new forms of knowledge representation. This knowledge is expressed by the PSNR matrix and the SSIM matrix, which represent information about the network inference process, i.e. the flow defined in this work. We verify the superiority, compatibility, and feasibility of FLRKD through three different sets of experiments. The experimental results show that FLRKD combined with the KD algorithm is superior to the most advanced relationship-based knowledge distillation method. In addition, there are still some problems to be solved in this work, such as how to select a more reasonable location for the hidden layer of the feature map extraction. If we put the perspective on the whole lightweight work, how to combine FLRKD with pruning or other lightweight methods, etc., all these are the problems we need to further think about and solve in the follow-up work.

References

1 A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.

2 Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distil- lation: Fast optimization, network minimization and transfer learning. In 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

3 Chengpeng Chen, Zichao Guo, Haien Zeng, Pengfei Xiong, and Jian Dong. Repghost: A hardware-efficient ghost module via re-parameterization. *arXiv preprint arXiv:2211.06088*, 2022.

4 Jie Song, Haofei Zhang, Xinchao Wang, Mengqi Xue, Ying Chen, Li Sun, Dacheng Tao, and Mingli Song. Tree-like decision distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13488–13497, 2021.

5 Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowl- edge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.