Knowledge Distillation Beyond Model Compression

Fahad Sarfraz*, Elahe Arani*, Bahram Zonooz
Advanced Research Lab, NavInfo Europe, Eindhoven, The Netherlands

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Abstract

Knowledge distillation (KD) is commonly deemed as an effective model compression technique in which a compact model (student) is trained under the supervision of a larger pretrained model or an ensemble of models (teacher).

Despite the recent advances, a clear understanding of where knowledge resides in a deep neural network and an optimal method for capturing knowledge from teacher and transferring it to student remains an open question.

Here, we provide an extensive study on nine different KD methods which covers a broad spectrum of approaches to capture and transfer knowledge. We further show the effectiveness of the KD framework in learning efficiently under varying severity levels of label noise and class imbalance. We demonstrate that the efficacy of KD goes much beyond a model compression technique and it should be considered as a general-purpose training paradigm which offers more robustness to common challenges in the real-world dataset.

Background

Effective deployment in the real-world necessitates developing compact networks that generalize well. To this end, several model compression techniques have been proposed [1]. Our study focuses on KD as an interactive learning framework which is more similar to how humans learn and provides a training paradigm instead of a compression technique.

Since the original formulation by Hinton [2], several distillation methods have been proposed. However, the effectiveness of the approach is dependent upon a number of factors: the capacity gap, the nature and degree of the constraint put on student training, and the characteristic of the teacher mimicked.

Hence, it is important to extensively study the effectiveness and versatility of different KD methods under a uniform experimental setting to gain further insights

Furthermore, we hypothesize that the additional supervision about the structural similarities between classes and/or data samples in KD can overcome many of the shortcomings of standard training procedure. We simulate varying degrees of label noise and class imbalance and demonstrate the robustness of KD to these common challenges.

Goal

The aim of the study is manifold:

i. provide extensive analysis of how the underlying mechanisms of different KD methods affect the generalization performance of the student.
ii. Demonstrate the versatility of the KD framework.
iii. Highlight the efficacy of KD framework as a general-purpose training framework which provides additional benefits over model compression.

Knowledge Distillation Methods

Here, we cover a diverse set of KD methods which differ from each other with respect to how knowledge is defined and transferred from the teacher.

a) ***Response Distillation:*** aims to mimic the output of the teacher. The key idea is that student can be trained to generalize the same way as the teacher by using the output probabilities produced by the teacher as a "soft target". **Hinton** [2] minimize the KL divergence between the smoother output probabilities and **BSS**[3] explicitly matches the decision boundary by utilizing an adversarial attack to discover samples supporting a decision boundary.

b) ***Representation Space Distillation:*** aims to mimic the latent feature space of the teacher. **FitNet**[4] uses intermediate-level hints from the teacher's hidden layers. **FSP**[5] eases the constraints and instead captures the transformation of features between the layers. **AT**[6] uses attention as a mechanism of transferring knowledge.

c) ***Relational Knowledge Distillation:*** aims to mimic the structural relations between the learned representation of the teacher using the mutual relations of data samples in the teacher's output representation. **RKD**[7] trains the student to form the same relational structure with that of the teacher. **SP**[8] encourages the student to preserve the pairwise similarities in the teacher.

d) ***Online Knowledge Distillation:*** updates both the student and teacher simultaneously. **DML**[9] involves knowledge sharing between a cohort of compact models trained collaboratively. **ONE**[10] uses a single multi-branch network and uses an ensemble of the branches as a stronger teacher to assist the learning of the target network.

A. Generalization Performance and Key Insights

KD aims to minimize the generalization gap between the teacher and the student. Despite the performance gains, there is still a considerable performance gap between student and teacher. A number of methods have been proposed to decrease this gap which differ from each other with respect to how knowledge is defined and transferred from the teacher. To highlight the subtle differences among the distillation methods used in the study, we present a broad categorization of these methods.

Therefore, the generalization gain over the baseline (a model trained without teacher supervision) is a key metric for evaluating the effectiveness of a KD method. Tables 1 demonstrate the effectiveness and versatility of the different KD methods in improving the generalization performance of the student on CIFAR-10. We also evaluated on the more complicated CIFAR-100 dataset and observe similar trends. For detailed analysis of the results please refer to the paper.

Key Insights:
From the empirical study, we derive the following insights, which can provide some guidelines for designing effective KD methods:

1. KD is an effective and versatile technique which consistently provides generalization gains on different datasets and network architectures even for the higher capacity gap between the student and teacher.
2. Generally, we observe that the methods which provide more flexibility to the student in learning, e.g. response distillation and relational KD methods are more versatile and can provide higher performance gains.
3. The performance of relational knowledge distillation methods provides a compelling case for the effectiveness of using the relations of the learned representations for KD. Furthermore, angular information can capture higher-level structure which aids in a performance gain.
4. Online distillation is a promising direction which removes the necessity of having a large pre-trained teacher for supervision and instead relies on mutual learning between a cohort of student models collectively supervising each other. This highlights the effectiveness of collaborative learning in improving the generalization of the models.

Table 1. Test set performance (%) on CIFAR-10. The best results are in bold. We run each experiment for 5 different seeds and report the mean ± 1 STD.

	ResNet-8	ResNet-14	ResNet-20	ResNet-26	WRN-10-2	WRN-16-2	WRN-28-2	WRN-40-2
Baseline	87.64±0.25	91.44±0.15	92.64±0.18	93.32±0.37	90.62±0.15	93.95±0.18	94.82±0.10	95.01±0.11
Hinton	88.80±0.16	**92.50±0.19**	93.25±0.18	93.58±0.10	91.72±0.12	94.28±0.09	94.97±0.10	95.12±0.10
BSS	89.18±0.43	91.99±0.20	92.92±0.18	93.52±0.08	**92.32±0.21**	94.27±0.18	94.72±0.15	94.96±0.20
FitNet	88.89±0.21	**92.50±0.10**	93.27±0.15	93.58±0.10	91.65±0.08	94.34±0.11	94.94±0.14	95.10±0.14
FSP	88.77±0.41	92.18±0.19	93.29±0.30	93.73±0.16	91.70±0.26	94.31±0.08	95.06±0.19	95.15±0.19
AT	86.07±0.32	91.66±0.16	92.96±0.09	93.32±0.14	90.99±0.21	94.50±0.18	**95.32±0.20**	95.39±0.15
SP	86.62±0.26	92.34±0.19	93.28±0.07	93.70±0.23	91.27±0.26	**94.64±0.17**	95.25±0.14	95.35±0.11
RKD-D	87.48±0.21	91.87±0.19	92.94±0.30	93.56±0.16	90.99±0.17	94.42±0.15	95.09±0.08	95.31±0.13
RKD-A	87.32±0.24	92.01±0.14	**93.30±0.12**	93.67±0.13	90.98±0.31	94.62±0.14	95.23±0.13	95.36±0.27
RKD-DA	87.14±0.19	92.05±0.20	93.05±0.20	93.73±0.09	90.92±0.16	94.52±0.11	95.19±0.12	**95.41±0.07**
ONE	**89.54±0.17**	92.30±0.23	93.27±0.16	**93.80±0.13**	87.75±1.92	92.80±0.08	94.70±0.18	95.11±0.09
DML	87.94±0.15	92.20±0.18	93.14±0.06	93.45±0.10	91.60±0.28	94.38±0.15	95.17±0.10	95.33±0.09

B. Label Noise

Learning efficiently under label noise is a major challenge [11] and one reason for the failure of standard training is that the only supervision the model receives is the one-hot-labels. KD on the other hand, can provides additional supervision e.g., the relative probabilities amongst the classes.

We hypothesis that the extra supervision signals in the KD framework can mitigate the adverse effect of incorrect ground truth labels. For evaluation, we simulate uniform label corruption on CIFAR-10. Table 2 shows that majority of the KD methods improve the generalization of the student trained under varying degrees of label corruption over the baseline.

Table 2. Test set performance (%) on CIFAR-10 with different label noise rates, σ. The best results are in bold, and the results below the baseline are colored in blue.

σ	0	0.2	0.4	0.6
Baseline	93.95±0.18	79.44±0.29	64.47±1.06	47.84±1.81
Hinton	94.28±0.09	**87.23±0.26**	76.32±0.87	58.18±0.35
BSS	94.27±0.18	80.28±0.33	71.46±0.20	47.69±0.37
FitNet	94.34±0.11	87.01±0.27	**76.73±0.52**	58.12±1.00
FSP	94.31±0.08	87.14±0.38	76.47±0.24	58.07±0.55
AT	94.50±0.18	79.59±0.47	64.46±0.88	46.44±0.78
SP	**94.64±0.17**	83.77±0.61	70.32±0.76	49.46±0.57
RKD-D	94.42±0.15	79.94±0.59	64.05±0.47	48.37±1.62
RKD-A	94.62±0.14	80.26±0.33	64.61±1.04	47.94±1.14
RKD-DA	94.52±0.11	80.45±0.58	65.10±1.08	48.90±0.52
ONE	92.80±0.08	83.76±0.40	68.64±0.53	40.49±1.12
DML	94.38±0.15	85.63±0.33	76.33±0.32	**59.89±1.66**

C. Class Imbalance

High class imbalance biases the models towards the prevalent classes [12]. In standard training, the model does not receive any information about the similarities between data points of different classes which can be useful in learning better representation for the minority classes.

We hypothesize that the additional relational information in KD, e.g relative probabilities or pairwise similarities, can be useful in learning the minority classes better. We simulate varying degrees of class imbalance using the power law as in [13] and demonstrate the effectiveness of KD.

Table 3. Test set performance (%) on CIFAR-10 with different class imbalance rates, γ. The best results are in bold, and the results below the baseline are colored in blue.

γ	0.20	0.60	1	2
Baseline	78.05±0.58	78.83±0.41	80.09±0.38	83.33±0.24
Hinton	79.15±0.28	80.08±0.25	81.18±0.51	83.69±0.69
BSS	78.07±0.20	79.22±0.53	80.44±0.24	82.15±0.22
FitNet	79.14±0.28	80.07±0.37	81.15±0.32	83.55±0.32
FSP	79.26±0.43	80.03±0.50	81.12±0.43	83.60±0.25
AT	79.13±0.40	80.51±0.23	80.96±0.18	84.13±0.32
SP	78.21±0.73	79.44±0.29	80.33±0.50	83.08±0.29
RKD-D	79.12±0.26	80.57±0.45	81.48±0.57	84.13±0.42
RKD-A	**79.52±0.51**	80.54±0.17	**81.52±0.36**	**84.33±0.42**
RKD-DA	79.43±0.41	**80.63±0.20**	81.50±0.37	84.02±0.21
ONE	77.48±1.05	78.04±0.86	79.48±0.39	80.88±1.05
DML	78.99±0.33	80.34±0.66	81.33±0.31	84.06±0.42

Conclusion

Our study emphasizes that knowledge distillation should not only be considered as an efficient model compression technique but rather as a general-purpose training paradigm that offers more robustness to common challenges in the real-world datasets compared to the standard training procedure.

Contact Information

Fahad.sarfraz@navinfo.eu
Elahe.arani@navinfo.eu
Bahram.zonooz@gmail.com

*** Equal Contribution**

References

1. C.Yu, et al. "A survey of model compression and acceleration for deep neural networks." *arXiv:1710.09282* (2017).
2. G. Hinton, et al, "Distilling the knowledge in a neural network," NeurIPS, 2014, Deep Learning Workshop.
3. B. Heo, et al, "Knowledge distillation with adversarial samples supporting decision boundary," AAAI, 2019
4. A. Romero, et al, "Fitnets: Hints for thin deep nets," ICLR, 2015.
5. J. Yim, et al, "A gift from knowledge distillation," CVPR, 2017.
6. S. Zagoruyko and N. Komodakis, "Paying more attention to attention," ICLR, 2017.
7. W. Park, et al, "Relational knowledge distillation," CVPR, 2019.
8. F. Tung and G. Mori, "Similarity-preserving knowledge distillation," ICCV, 2019.
9. Y. Zhang, et al, "Deep mutual learning," CVPR, 2018.
10. X. Lan, et al, "Knowledge distillation by on-the-fly native ensemble," NeurIPS, 2018.
11. S. Sukhbaatar, et al, "Training convolutional networks with noisy labels," arXiv:1406.2080, 2014
12. G. Van Horn, et al, "The inaturalist species classification and detection dataset," CVPR 2018
13. Q. Dong, et al, "Imbalanced deep learning by minority class incremental rectification," TPAMI 2018.