25th International Conference on
Pattern Recognition

Knowledge Distillation Beyond Model Compression

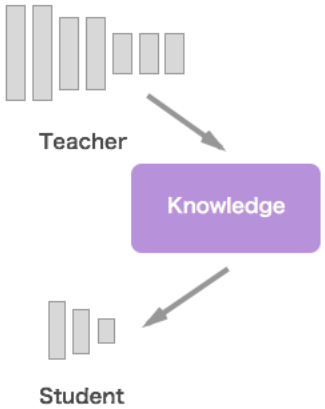


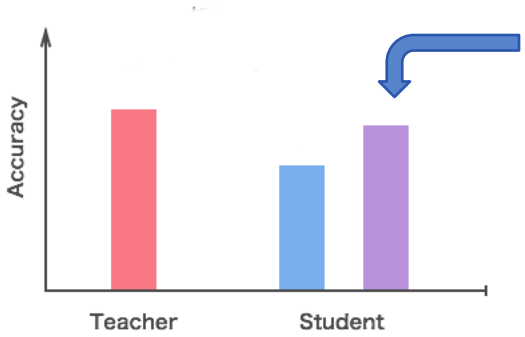
Fahad Sarfraz*, Elahe Arani*, Bahram Zonooz

NAVINFO
四 维 图 新
Europe
Advanced Research Lab

Knowledge Distillation

Knowledge distillation (KD) involves training a compact network (student) under the supervision of a larger pre-trained network or an ensemble of models (teacher) in an interactive manner which is more similar to how humans learn.





KD leads to consistent performance gain compared to the student model trained alone without teacher supervision.

https://medium.com/neuralmachine/knowledge-distillation-dc241d7c2322

Goal of the Study

Extensive study of KD methods

- Study the effectiveness and versatility of different KD methods capturing different aspects of the teacher.

General-purpose training framework

- Analyze the characteristics of KD as a general-purpose training framework beyond just model compression.

Generalization Gap

A clear understanding of where knowledge resides in a deep neural network is still lacking and consequently an optimal method of capturing knowledge from teacher and transferring it to student remains an open question.

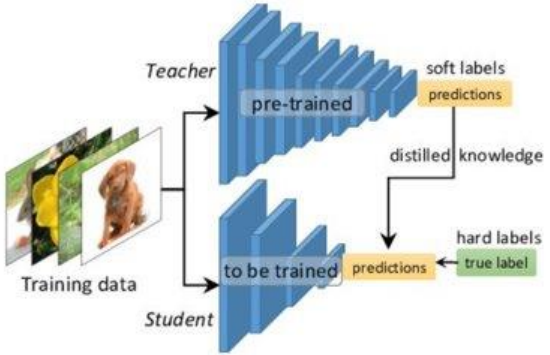
- Despite the performance gains, there is still a considerable performance gap between student and teacher.
- A number of methods have been proposed to decrease this gap which differ from each other with respect to how knowledge is defined and transferred from the teacher.
- To highlight the subtle differences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

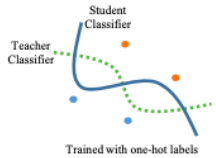
Response Distillation

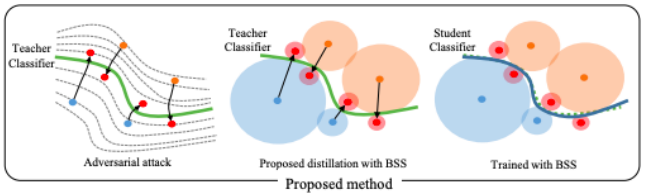
Aims to mimic the output of the teacher. It can be seen as an implicit method for matching the decision boundaries of the student and the teacher.

Methods:

- **Hinton:** proposed to raise the temperature of the final softmax function and minimize the Kullback–Leibler (KL) divergence between the smoother output probabilities.
- **Boundary Support Distillation (BSS):** explicitly matches the decision boundary by utilizing an adversarial attack to discover samples supporting a decision boundary.







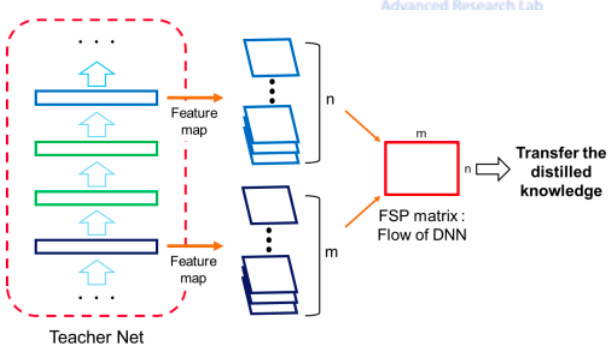
G. Hinton, et al, “Distilling the knowledge in a neural network,” NeurIPS, 2014, Deep Learning Workshop.
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764
B. Heo, et al, “Knowledge distillation with adversarial samples supporting decision boundary,” AAAI, 2019

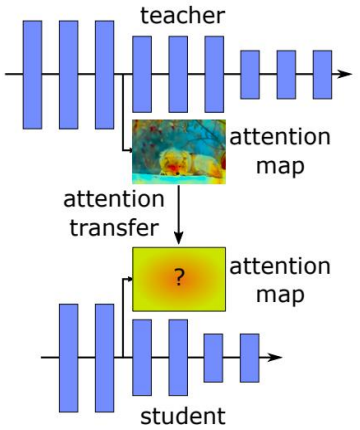
Representation Space Distillation

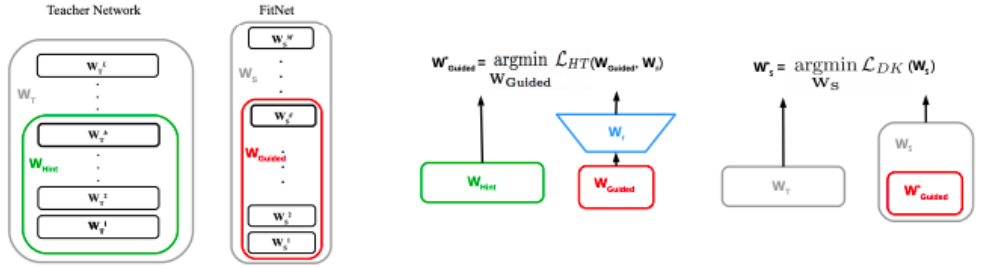
Aims to mimic the latent feature space of the teacher.

Methods:

- **FitNet** uses intermediate-level hints from the teacher's hidden layers and puts a hard constraint on student to exactly match the representation space of teacher.
- **FSP** eases the constraints and instead captures the transformation of features between the layers.
- **AT** uses attention as a mechanism of transferring knowledge.







A. Romero, et al, "Fitnets: Hints for thin deep nets," ICLR, 2015.
J. Yim, et al, "A gift from knowledge distillation," CVPR, 2017.
S. Zagoruyko and N. Komodakis, "Paying more attention to attention," ICLR, 2017.

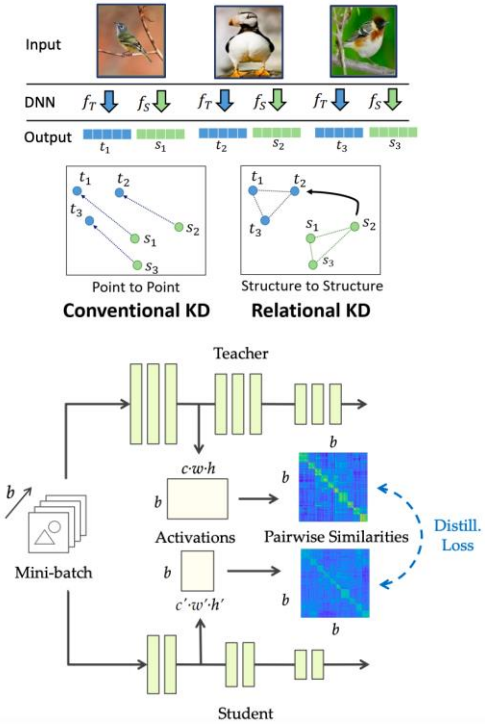
Relational Knowledge Distillation

Mimic the structural relations between the learned representation of the teacher using the mutual relations of data samples in the teacher's output representation.

Methods:

- **RKD** trains the student to form the same relational structure with that of the teacher in terms of two variants of relational potential functions: Distance-wise potential, **RKD-D**, and angle wise potential, **RKD-A**. **RKD-DA** combines both of these losses to train the student.

- **SP** encourages the student to preserve the pairwise similarities in the teacher in such a way that data pairs that produce similar/dissimilar activations in the teacher, also produce similar/dissimilar activations in the student.



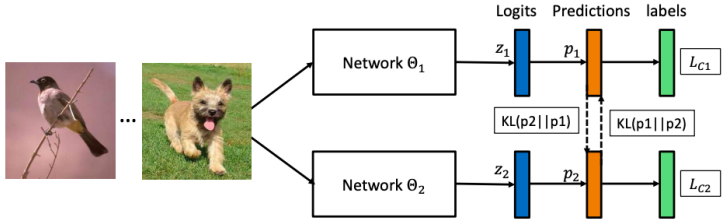
W. Park, et al, "Relational knowledge distillation," CVPR, 2019.
F. Tung and G. Mori, "Similarity-preserving knowledge distillation," ICCV, 2019.

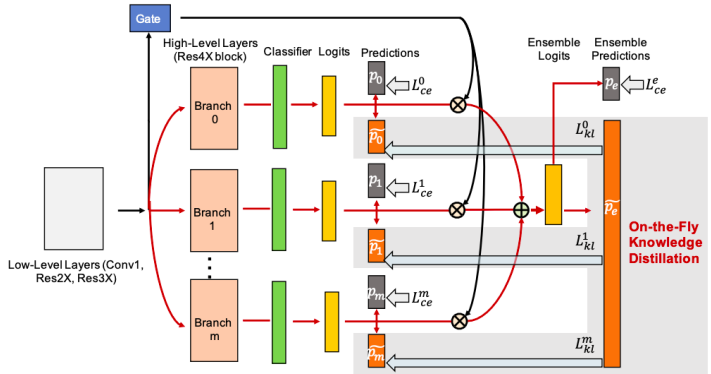
Online Knowledge Distillation

Aims to circumvent the need for a static teacher and updates both the student and teacher simultaneously.

Methods:

- **DML** involves knowledge sharing between a cohort of compact models trained collaboratively.
- **ONE** 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.





Y. Zhang, et al, "Deep mutual learning," CVPR, 2018.

X. Lan, et al, "Knowledge distillation by on-the-fly native ensemble," NeurIPS, 2018.

Generalization Performance: CIFAR-10

TABLE III: 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

Generalization Performance: CIFAR-100

TABLE IV: Test set performance (%) on CIFAR-100. 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	71.78±0.26	76.95±0.43	77.92±0.40	78.82±0.24	67.99±0.55	72.35±0.36	74.93±0.39	75.94±0.12
Hinton	72.78±0.36	78.18±0.14	79.58±0.22	79.73±0.30	67.70±0.54	74.12±0.37	76.08±0.36	77.07±0.13
BSS	73.02±0.07	76.96±0.19	78.37±0.27	78.66±0.23	**69.55±0.38**	73.04±0.21	75.59±0.24	76.55±0.16
FitNet	72.86±0.28	78.48±0.30	79.55±0.18	79.83±0.34	67.83±0.47	72.82±2.40	76.31±0.09	77.25±0.15
FSP	72.93±0.24	78.34±0.45	79.65±0.18	79.62±0.18	67.64±0.19	73.86±0.27	76.21±0.14	77.09±0.27
AT	71.99±0.08	76.88±0.20	78.35±0.10	78.94±0.34	67.45±0.27	72.78±0.32	75.51±0.13	76.60±0.13
SP	73.18±0.24	**78.53±0.29**	**79.76±0.27**	**79.93±0.29**	66.77±0.27	73.42±0.37	76.52±0.35	77.43±0.14
RKD-D	71.99±0.23	77.02±0.21	78.23±0.26	78.80±0.28	68.14±0.34	72.52±0.30	75.48±0.33	76.34±0.29
RKD-A	71.95±0.33	76.93±0.35	78.51±0.25	79.10±0.18	68.10±0.31	72.87±0.23	75.50±0.41	76.97±0.17
RKD-DA	71.70±0.19	77.14±0.40	78.64±0.21	79.16±0.11	67.94±0.37	72.88±0.23	75.73±0.32	76.91±0.22
ONE	73.30±0.12	78.04±0.07	79.24±0.18	79.74±0.27	57.38±2.11	69.78±0.94	74.49±0.54	76.89±0.27
DML	**73.57±0.09**	78.07±0.20	79.15±0.22	79.32±0.38	68.99±0.23	**74.44±0.25**	**76.65±0.17**	**77.65±0.19**

Generalization Performance: Key Findings

- KD is an effective and versatile technique which consistently provides generalization gains even when the capacity gap is large.
- Generally, we observe that the methods which provide more flexibility to the student in learning are more versatile and can provide higher performance gains.
- 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 performance gain.
- Online distillation is a promising direction which highlights the effectiveness of collaborative learning in improving the generalization of the models.

General-Purpose Training Framework: Label Noise

- A major reason for the failure of standard training is that the only supervision the model receives is the one-hot-labels.
- In KD on the other hand, in addition to the ground truth label, the model receives supervision from the teacher, e.g. the soft probabilities, relational knowledge or the consensus between different students.

We hypothesize that the extra supervision signals in the KD framework can mitigate the adverse effect of incorrect ground truth labels.

General-Purpose Training Framework: Label Noise

TABLE V: 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. We run each experiment for 5 different seeds and report the mean $\pm$ 1 STD.

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**

General-Purpose Training Framework: Class Imbalance

- Models trained with standard training exhibit bias towards the prevalent classes at the expense of minority classes.
- Because of the one-hot-encoded labels in standard training, the only information the model receives about a particular class comes from the datapoints belonging to it.
- 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 can be useful in learning the minority classes better.

General-Purpose Training Framework: Class Imbalance

TABLE VI: 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. We run each experiment for 5 different seeds and report the mean $\pm$ 1 STD.

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

Takeaway

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.

THANK YOU