Decoupled Knowledge Distillation (DKD)

Borui Zhao[1] Quan Cui[2] Renjie Song[1] Yiyu Qiu[1,3] Jiajun Liang[1]

[1]MEGVII Technology [2]Waseda University [3]Tsinghua University

zhaoborui.gm@gmail.com, cui-quan@toki.waseda.jp,

chouyy18@mails.tsinghua.edu.cn, {songrenjie, liangjiajun}@megvii.com

2023. 02. 08

Presenter: Seonghak KIM

Introduction

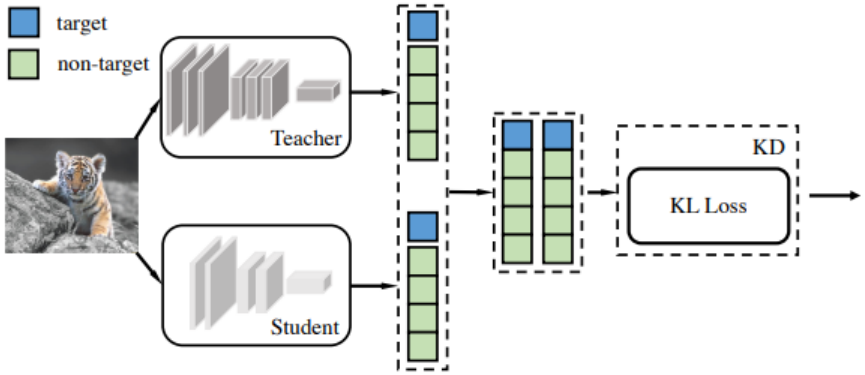
- **Knowledge Distillation**
  - **Logits-based method**
    - (+) computational and storage cost ↓
    - (−) unsatisfactory performance
  - **Feature-based method**
    - (+) superior performance
    - (−) extra computational cost and storage usage

→ *Potential of logit distillation is limited.*

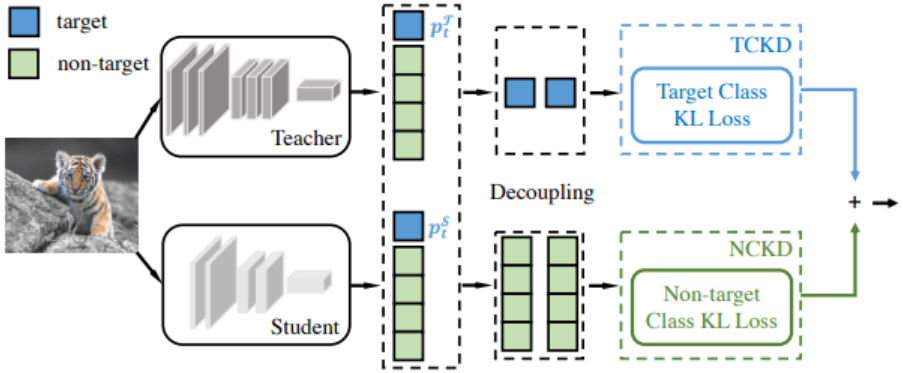
- **Decoupled Knowledge Distillation (DKD)**
  - **Target classification knowledge distillation (TCKD)**
    - Binary logit distillation
  - **Non-target classification knowledge distillation (NCKD)**
    - Knowledge among non-target logits

Introduction

- **Decoupled Knowledge Distillation (DKD)**



(a) Classical Knowledge Distillation (KD).



$$\textbf{\textit{Classical KD}} = TCKD + (1 - p_t^T) * NCKD$$

$$\textbf{\textit{DKD(Ours)}} = \alpha * TCKD + \beta * NCKD$$

(b) Decoupled Knowledge Distillation (DKD).

[Illustration of the classical KD and DKD]

Reformulation

Notations

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}, \qquad \boldsymbol{p} = [p_1, p_2, \dots, p_t, \dots, p_C] \in \mathbb{R}^{1\times C}$$

- Binary probabilities

$$p_t = \frac{\exp(z_t)}{\sum_{j=1}^{C} \exp(z_j)}, p_{\setminus t} = \frac{\sum_{k=1,k\neq t}^{C} \exp(z_k)}{\sum_{j=1}^{C} \exp(z_j)}, \qquad \boldsymbol{b} = [p_t, p_{\setminus t}] \in \mathbb{R}^{1\times 2}$$

- Probabilities among non-target classes

$$\hat{p}_i = \frac{\exp(z_i)}{\sum_{j=1,j\neq t}^{C} \exp(z_j)}, \qquad \hat{\boldsymbol{p}} = [\hat{p}_1, \dots, \hat{p}_{t-1}, \hat{p}_{t+1}, \dots, \hat{p}_C] \in \mathbb{R}^{1\times(C-1)}$$

Method

- **Reformulation**
  - **Vanilla KD**

$$\begin{aligned}
\mathrm{KD} &= \mathrm{KL}(\mathbf{p}^{\mathcal{T}}||\mathbf{p}^{\mathcal{S}}) \\
&= p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + \sum_{i=1, i\neq t}^{C} p_i^{\mathcal{T}} \log(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}}).
\end{aligned}$$

$$\begin{aligned}
\mathrm{KD} &= p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + p_{\backslash t}^{\mathcal{T}} \sum_{i=1, i\neq t}^{C} \hat{p}_i^{\mathcal{T}} (\log(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}}) + \log(\frac{p_{\backslash t}^{\mathcal{T}}}{p_{\backslash t}^{\mathcal{S}}})) \\
&= \underbrace{p_t^{\mathcal{T}} \log(\frac{p_t^{\mathcal{T}}}{p_t^{\mathcal{S}}}) + p_{\backslash t}^{\mathcal{T}} \log(\frac{p_{\backslash t}^{\mathcal{T}}}{p_{\backslash t}^{\mathcal{S}}})}_{\mathrm{KL}(\mathbf{b}^{\mathcal{T}}||\mathbf{b}^{\mathcal{S}})} + \underbrace{p_{\backslash t}^{\mathcal{T}} \sum_{i=1, i\neq t}^{C} \hat{p}_i^{\mathcal{T}} \log(\frac{\hat{p}_i^{\mathcal{T}}}{\hat{p}_i^{\mathcal{S}}})}_{\mathrm{KL}(\hat{\mathbf{p}}^{\mathcal{T}}||\hat{\mathbf{p}}^{\mathcal{S}})}
\end{aligned}$$

$$\mathrm{KD} = \mathrm{KL}(\mathbf{b}^{\mathcal{T}}||\mathbf{b}^{\mathcal{S}}) + (1 - p_t^{\mathcal{T}})\mathrm{KL}(\hat{\mathbf{p}}^{\mathcal{T}}||\hat{\mathbf{p}}^{\mathcal{S}})$$

$$\boxed{\mathrm{KD} = \mathrm{TCKD} + (1 - p_t^{\mathcal{T}})\mathrm{NCKD}.}$$

- While NCKD focuses on the knowledge among non-target classes, TCKD focus on the knowledge related to the target class.

Method

Effects of TCKD and NCKD

student	TCKD	NCKD	top-1	Δ
ResNet32×4 as the teacher				
ResNet8×4			72.50	-
	✓	✓	73.63	+1.13
	✓		68.63	-3.87
		✓	74.26	+1.76
ShuffleNet-V1			70.50	-
	✓	✓	74.29	+3.79
	✓		70.52	+0.02
		✓	74.91	+4.41
WRN-40-2 as the teacher				
WRN-16-2			73.26	-
	✓	✓	74.96	+1.70
	✓		70.96	-2.30
		✓	74.76	+1.50
ShuffleNet-V1			70.50	-
	✓	✓	74.92	+4.42
	✓		70.62	+0.12
		✓	75.12	+4.62

Table 1. Accuracy(%) on the CIFAR-100 validation set. Δ represents the performance improvement over the baseline.

- Singly applying TCKD is unhelpful or even harmful. (−)
- Performance of NCKD are comparable and even better than vanilla KD (−)

∴ **target-class-related knowledge could not be as important as knowledge among non-target classes.**

Effects of TCKD and NCKD

student	TCKD	top-1	Δ
ResNet8×4		73.82	-
	✓	75.33	+1.51
ShuffleNet-V1		77.13	-
	✓	77.98	+0.85

Table 2. Accuracy(%) on the CIFAR-100 validation. We set ResNet32×4 as the teacher and ResNet8×4 as the student. Both teachers and students are trained with AutoAugment [5].

noisy ratio	TCKD	top-1	Δ
0.1		70.99	-
	✓	70.96	-0.03
0.2		67.55	-
	✓	68.03	+0.48
0.3		64.62	-
	✓	65.26	+0.64

Table 3. Accuracy(%) on the CIFAR-100 validation with different noisy ratios on the training set. We set ResNet32×4 as the teacher and ResNet8×4 as the student.

TCKD	top-1	Δ
	70.71	-
✓	71.03	+0.32

Table 4. Accuracy(%) on the ImageNet validation. We set ResNet-34 as the teacher and ResNet-18 as the student.

∴ The more difficult the training data is, the more benefits TCKD could provide.

Method

Decoupled Knowledge Distillation (DKD)

$$\mathrm{KD} = \mathrm{TCKD} + (1 - p_t^{\mathcal{T}})\mathrm{NCKD}.$$

- NCKD loss is coupled with $\left(1 - p_t^{\mathcal{T}}\right)$.

→ More confident predictions results in smaller NCKD weights. (highly suppressed weights)

- Weights of NCKD and TCKD are coupled.

$$\therefore \mathrm{DKD} = \alpha\mathrm{TCKD} + \beta\mathrm{NCKD}$$

where α and β are hyper-parameters.

Experiments

- **Ablation: α and β**
  - Teacher: ResNet32×4, Student: ResNet8× 4

β	$1-p_t^{\mathcal{T}}$	1.0	2.0	4.0	8.0	10.0
top-1	73.63	74.79	75.44	75.94	**76.32**	76.18
α	0.0	0.2	0.5	1.0	2.0	4.0
top-1	75.30	75.64	76.12	**76.32**	76.11	75.42

- **CIFAR-100**

distillation manner	teacher	ResNet56	ResNet110	ResNet32×4	WRN-40-2	WRN-40-2	VGG13
		72.34	74.31	79.42	75.61	75.61	74.64
	student	ResNet20	ResNet32	ResNet8×4	WRN-16-2	WRN-40-1	VGG8
		69.06	71.14	72.50	73.26	71.98	70.36
features	FitNet [28]	69.21	71.06	73.50	73.58	72.24	71.02
	RKD [23]	69.61	71.82	71.90	73.35	72.22	71.48
	CRD [33]	71.16	73.48	75.51	75.48	74.14	73.94
	OFD [10]	70.98	73.23	74.95	75.24	74.33	73.95
	ReviewKD [1]	71.89	73.89	75.63	76.12	**75.09**	**74.84**
logits	KD [12]	70.66	73.08	73.33	74.92	73.54	72.98
	DKD	**71.97**	**74.11**	**76.32**	**76.24**	74.81	74.68
	Δ	+1.31	+1.03	+2.99	+1.32	+1.27	+1.70

Table 6. **Results on the CIFAR-100 validation.** Teachers and students are in the **same** architectures. And Δ represents the performance improvement over the classical KD. All results are the average over 5 trials.

Experiments

CIFAR-100

distillation manner	teacher / student	ResNet32×4 79.42 / ShuffleNet-V1 70.50	WRN-40-2 75.61 / ShuffleNet-V1 70.50	VGG13 74.64 / MobileNet-V2 64.60	ResNet50 79.34 / MobileNet-V2 64.60	ResNet32×4 79.42 / ShuffleNet-V2 71.82
features	FitNet [28]	73.59	73.73	64.14	63.16	73.54
	RKD [23]	72.28	72.21	64.52	64.43	73.21
	CRD [33]	75.11	76.05	69.73	69.11	75.65
	OFD [10]	75.98	75.85	69.48	69.04	76.82
	ReviewKD [1]	**77.45**	**77.14**	**70.37**	69.89	**77.78**
logits	KD [12]	74.07	74.83	67.37	67.35	74.45
	DKD	76.45	76.70	69.71	**70.35**	77.07
	Δ	+2.38	+1.87	+2.34	+3.00	+2.62

Table 7. **Results on the CIFAR-100 validation.** Teachers and students are in **different** architectures. And Δ represents the performance improvement over the classical KD. All results are the average over 5 trials.

ImageNet

distillation manner			features				logits		
	teacher	student	AT [43]	OFD [10]	CRD [33]	ReviewKD [1]	KD [12]	KD*	**DKD**
top-1	73.31	69.75	70.69	70.81	71.17	71.61	70.66	71.03	**71.70**
top-5	91.42	89.07	90.01	89.98	90.13	**90.51**	89.88	90.05	90.41

Table 8. **Top-1 and top-5 accuracy (%) on the ImageNet validation.** We set **ResNet-34** as the teacher and **ResNet-18** as the student. KD* represents the result of our implementation. All results are the average over 3 trials.

distillation manner			features				logits		
	teacher	student	AT [43]	OFD [10]	CRD [33]	ReviewKD [1]	KD [12]	KD*	**DKD**
top-1	76.16	68.87	69.56	71.25	71.37	**72.56**	68.58	70.50	72.05
top-5	92.86	88.76	89.33	90.34	90.41	91.00	88.98	89.80	**91.05**

Table 9. **Top-1 and top-5 accuracy (%) on the ImageNet validation.** We set **ResNet-50** as the teacher and **MobileNet-V2** as the student. KD* represents the result of our implementation. All results are the average over 3 trials.

Conclusions

- **Reformulation of vanilla KD loss into two parts**
  - **TCKD and NCKD**
- **Decoupled Knowledge Distillation**
  - **Coupled formulation limits the effectiveness of transfer**
- **Significant improvements on various datasets**

Thank you.