Korea university Biomedical Engineering

SNUH RADIOLOGY

CCIDS 의료영상데이터사이언스센터 Center for Clinical Imaging Data Science

2020. 05. 10

Regularizing Class-Wise Predictions via Self-knowledge Distillation

Sang Wook Kim
Korea university
Department of Biomedical Engineering

Contents

Introduction

1. Regularization is needed because DNNs suffer from **overfitting** and **poor generalizations**

2. Regularizing the **predictive distribution of DNNs** can be effective because **it contains the most succinct knowledge of the model**.

3. This paper focuses on developing **a new output regularizer** for DNN utilizing concept of **dark knowledge (the knowledge on wrong predictions made by DNN)**

Class-wise self-knowledge distillation

self-knowledge distillation ???

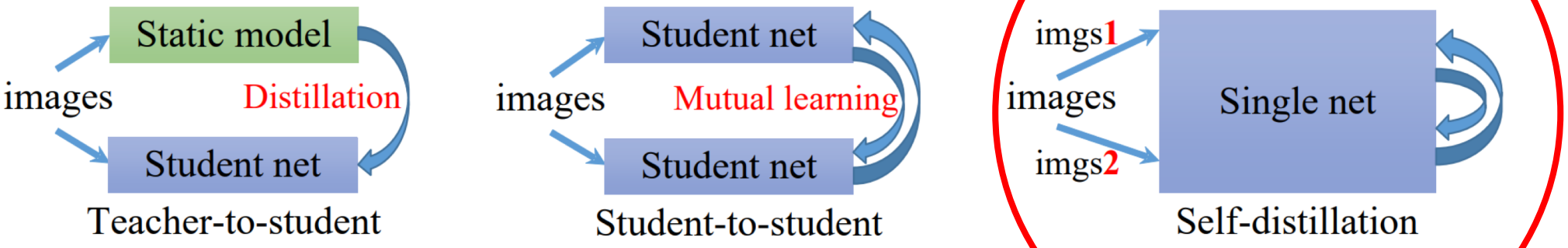


Figure 1: The diagrams of three distillation mechanisms.

Class-wise self-knowledge distillation (CS-KD)

1. Posterior predictive distribution (**Softmax** classifier)

$$P\left(y|\mathbf{x};\theta,T\right)=\frac{\exp\left(f_y\left(\mathbf{x};\theta\right)/T\right)}{\sum_{i=1}^{C}\exp\left(f_i\left(\mathbf{x};\theta\right)/T\right)},$$

2. Here, f_i denotes **the logit of DNNs** for class i which are **parameterized by θ**, and **T > 0 is the temperature scaling parameter**.

Class-wise regularization

1. Consider **matching predictive distributions** on samples of the same class

$\Rightarrow$ Propose a **class-wise regularization loss**

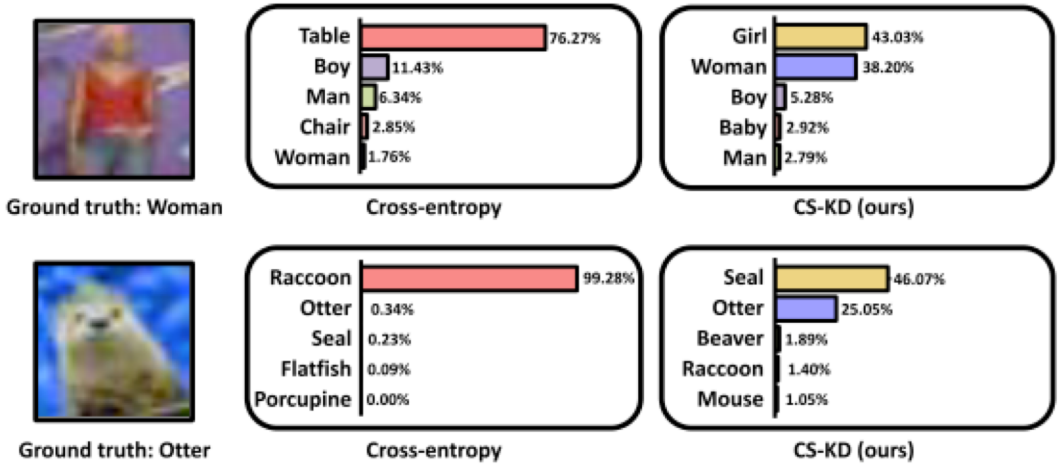
$\Rightarrow$ Enforces **consistent predictive distributions in the same class**

$$\mathcal{L}_{\texttt{cls}}\left(\mathbf{x}, \mathbf{x}'; \theta, T\right) := \mathrm{KL}\left(P(y|\mathbf{x}'; \widetilde{\theta}, T) \middle\| P(y|\mathbf{x}; \theta, T)\right)$$

$$\mathcal{L}_{\texttt{CS-KD}}(\mathbf{x}, \mathbf{x}', y; \theta, T) := \mathcal{L}_{\texttt{CE}}(\mathbf{x}, y; \theta) + \lambda_{\texttt{cls}} \cdot T^2 \cdot \mathcal{L}_{\texttt{cls}}(\mathbf{x}, \mathbf{x}'; \theta, T),$$

Effects of class-wise regularization

1. CS-KD is **the simplest way** to achieve below two goals
    1. **Preventing overconfident predictions**
    2. **Reducing the intra-class variations**



(b) Top-5 softmax scores on misclassified samples

(d) CS-KD (ours)

Solve Image Classification Problems

1. Data diversity => To demonstrate CS-KD under general situation
    1. CIFAR-100
    2. TinyImageNet
    3. CUB-200-2011
    4. MIT67

2. Network Architecture
    1. **ResNet**
    2. **DenseNet**

3. Hyperparameters
    1. **SGD** (0.9 momentum, 0.0001 weight decay)
    2. Standard **data augmentation** technique for ImageNet

Compare with **Prior Regularization methods**

1. AdaCos
2. Virtual-softmax
3. Maximum-entropy
4. Label-smoothing
5. DDGSD
6. BYOT

Evaluation Metric

1. Top-1 / 5 error rate

2. Expected Calibration Error (ECE)

3. Recall at k (R @ k)

Experiments & Results

Comparison with output regularization methods

Model	Method	CIFAR-100		TinyImageNet		CUB-200-2011		Stanford Dogs		MIT67	
ResNet-18	Cross-entropy	24.71$_{\pm 0.24}$		43.53$_{\pm 0.19}$		46.00$_{\pm 1.43}$		36.29$_{\pm 0.32}$		44.75$_{\pm 0.80}$	
	AdaCos	23.71$_{\pm 0.36}$		42.61$_{\pm 0.20}$		35.47$_{\pm 0.07}$		32.66$_{\pm 0.34}$		42.66$_{\pm 0.43}$	
	Virtual-softmax	23.01$_{\pm 0.42}$		42.41$_{\pm 0.20}$		35.03$_{\pm 0.51}$		31.48$_{\pm 0.16}$		42.86$_{\pm 0.71}$	
	Maximum-entropy	22.72$_{\pm 0.29}$		41.77$_{\pm 0.13}$		39.86$_{\pm 1.11}$		32.41$_{\pm 0.20}$		43.36$_{\pm 1.62}$	
	Label-smoothing	22.69$_{\pm 0.28}$		43.09$_{\pm 0.34}$		42.99$_{\pm 0.99}$		35.30$_{\pm 0.66}$		44.40$_{\pm 0.71}$	
	CS-KD (ours)	**21.99**$_{\pm 0.13}$	**(-11.0%)**	**41.62**$_{\pm 0.38}$	**(- 4.4%)**	**33.28**$_{\pm 0.99}$	**(-27.7%)**	**30.85**$_{\pm 0.28}$	**(-15.0%)**	**40.45**$_{\pm 0.45}$	**(- 9.6%)**
DeseNet-121	Cross-entropy	22.23$_{\pm 0.04}$		39.22$_{\pm 0.27}$		42.30$_{\pm 0.44}$		33.39$_{\pm 0.17}$		41.79$_{\pm 0.19}$	
	AdaCos	22.17$_{\pm 0.24}$		38.76$_{\pm 0.23}$		30.84$_{\pm 0.38}$		27.87$_{\pm 0.65}$		40.25$_{\pm 0.68}$	
	Virtual-softmax	23.66$_{\pm 0.10}$		41.58$_{\pm 1.58}$		33.85$_{\pm 0.75}$		30.55$_{\pm 0.72}$		43.66$_{\pm 0.30}$	
	Maximum-entropy	22.87$_{\pm 0.45}$		38.39$_{\pm 0.33}$		37.51$_{\pm 0.71}$		29.52$_{\pm 0.74}$		43.48$_{\pm 1.30}$	
	Label-smoothing	21.88$_{\pm 0.45}$		38.75$_{\pm 0.18}$		40.63$_{\pm 0.24}$		31.39$_{\pm 0.46}$		42.24$_{\pm 1.23}$	
	CS-KD (ours)	**21.69**$_{\pm 0.49}$	**(- 2.4%)**	**37.96**$_{\pm 0.09}$	**(- 3.2%)**	**30.83**$_{\pm 0.39}$	**(-27.1%)**	**27.81**$_{\pm 0.13}$	**(-16.7%)**	**40.02**$_{\pm 0.91}$	**(- 4.2%)**

Table 1. Top-1 error rates (%) on various image classification tasks and model architectures. We report the mean and standard deviation over three runs with different random seeds. Values in parentheses indicate relative error rate reductions from the cross-entropy, and the best results are indicated in bold.

Experiments & Results

Comparison with self-distillation methods

Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
Cross-entropy	$24.71_{\pm 0.24}$	$43.53_{\pm 0.19}$	$46.00_{\pm 1.43}$	$36.29_{\pm 0.32}$	$44.75_{\pm 0.80}$
DDGSD	$23.85_{\pm 1.57}$	**$41.48_{\pm 0.12}$**	$41.17_{\pm 1.28}$	$31.53_{\pm 0.54}$	$41.17_{\pm 2.46}$
BYOT	$23.81_{\pm 0.11}$	$44.02_{\pm 0.57}$	$40.76_{\pm 0.39}$	$34.02_{\pm 0.14}$	$44.88_{\pm 0.46}$
CS-KD (ours)	**$21.99_{\pm 0.13}$ (-11.0%)**	$41.62_{\pm 0.38}$ (- 4.4%)	**$33.28_{\pm 0.99}$ (-27.7%)**	**$30.85_{\pm 0.28}$ (-15.0%)**	**$40.45_{\pm 0.45}$ (- 9.6%)**

Table 2. Top-1 error rates (%) of ResNet-18 with self-distillation methods on various image classification tasks. We report the mean and standard deviation over three runs with different random seeds. Values in parentheses indicate relative error rate reductions from the cross-entropy, and the best results are indicated in bold. The self-distillation methods are re-implemented under our code-base.

Experiments & Results

Evaluation on Large-scale datasets

Model	Method	Top-1 (1-crop)
ResNet-50	Cross-entropy	24.0
	CS-KD (ours)	**23.6**
ResNet-101	Cross-entropy	22.4
	CS-KD (ours)	**22.0**
ResNeXt-101-32x4d	Cross-entropy	21.6
	CS-KD (ours)	**21.2**

Table 5. Top-1 error rates (%) on ImageNet dataset with various model architectures trained for 90 epochs with batch size 256. The best results are indicated in bold.

Experiments & Results

Compatibility with other regularization methods

Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
Cross-entropy	24.71 $\pm$0.24	43.53 $\pm$0.19	46.00 $\pm$1.43	36.29 $\pm$0.32	44.75 $\pm$0.80
CS-KD (ours)	21.99 $\pm$0.13	41.62 $\pm$0.38	33.28 $\pm$0.99	30.85 $\pm$0.28	40.45 $\pm$0.45
Mixup	21.67 $\pm$0.34	41.57 $\pm$0.38	37.09 $\pm$0.27	32.54 $\pm$0.04	41.67 $\pm$1.05
Mixup + CS-KD (ours)	**20.40** $\pm$**0.31**	**40.71** $\pm$**0.32**	**30.71** $\pm$**0.64**	**29.93** $\pm$**0.14**	**39.65** $\pm$**0.85**

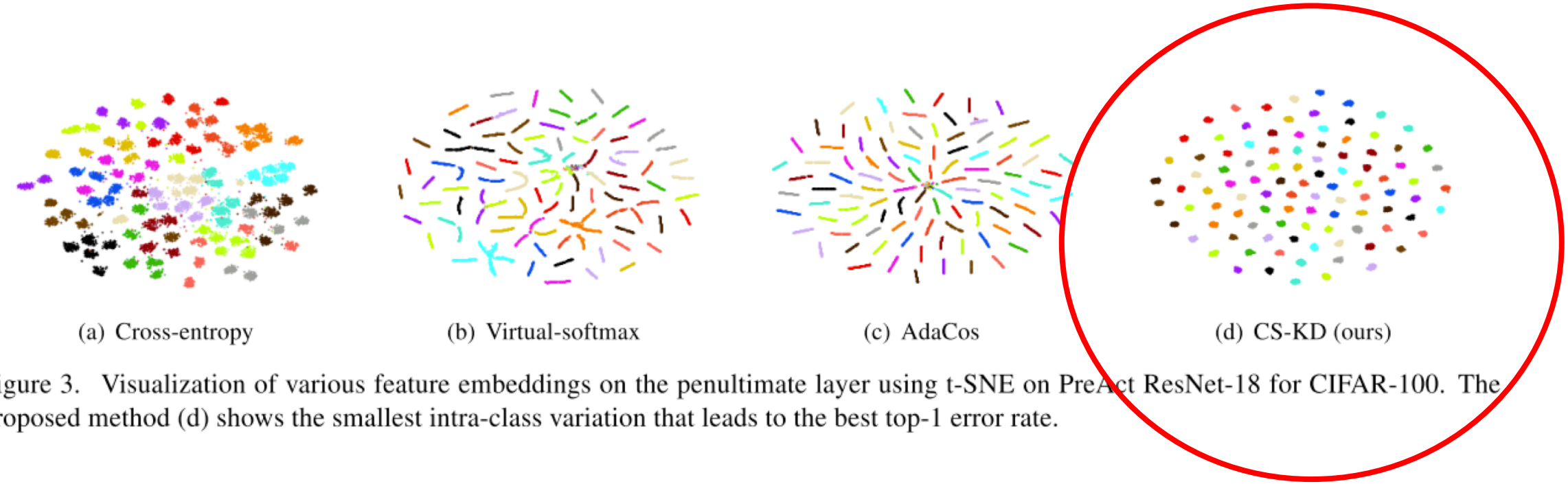
Table 3. Top-1 error rates (%) of ResNet-18 with Mixup regularization on various image classification tasks. We report the mean and standard deviation over three runs with different random seeds, and the best results are indicated in bold.

Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
Cross-entropy	26.72 $\pm$0.33	46.61 $\pm$0.22	48.36 $\pm$0.61	38.96 $\pm$0.40	44.75 $\pm$0.62
CS-KD (ours)	25.80 $\pm$0.10	44.67 $\pm$0.12	39.12 $\pm$0.09	34.07 $\pm$0.46	41.54 $\pm$0.67
KD	25.84 $\pm$0.07	43.31 $\pm$0.11	39.32 $\pm$0.65	34.23 $\pm$0.42	41.47 $\pm$0.79
KD + CS-KD (ours)	**25.58** $\pm$**0.16**	**42.82** $\pm$**0.33**	**34.47** $\pm$**0.17**	**32.59** $\pm$**0.50**	**40.27** $\pm$**0.78**

Table 4. Top-1 error rates (%) of ResNet-10 (student) with knowledge distillation (KD) on various image classification tasks. Teacher networks are pre-trained on DenseNet-121 by CS-KD. We report the mean and standard deviation over three runs with different random seeds, and the best results are indicated in bold.

Experiments & Results

Feature Embedding analysis



(a) Cross-entropy (b) Virtual-softmax (c) AdaCos (d) CS-KD (ours)

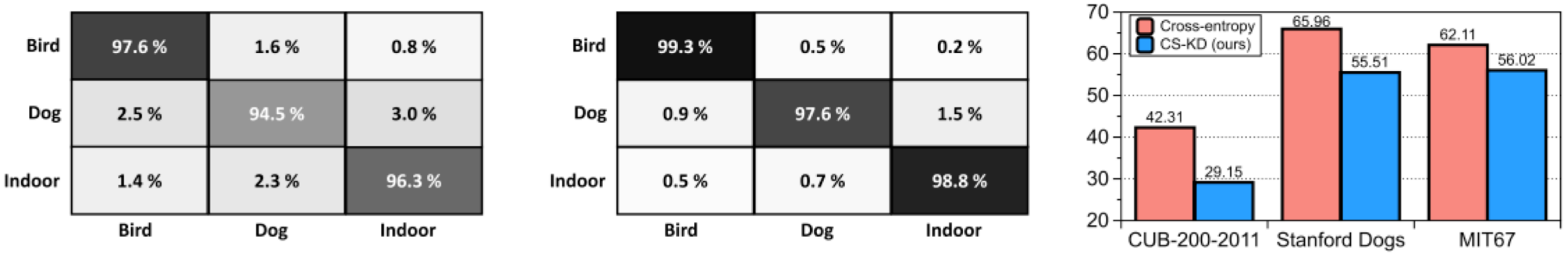
Figure 3. Visualization of various feature embeddings on the penultimate layer using t-SNE on PreAct ResNet-18 for CIFAR-100. The proposed method (d) shows the smallest intra-class variation that leads to the best top-1 error rate.

Experiments & Results

Hierarchical image classification

1. Increasing the **correlation between similar classes** in predictions

⇒ Expect the trained classifier capture a hierarchical structure



(a) Cross-entropy (b) CS-KD (ours) (c) Top-1 error rates (%)

Figure 4. Experimental results of ResNet-18 on the mixed dataset. The hierarchical classification accuracy (%) of each model trained by (a) the cross-entropy and (b) our method. One can observe that the model trained by CS-KD is less confusing classes across different domains. (c) Top-1 error rates (%) of fine-grained label classification.

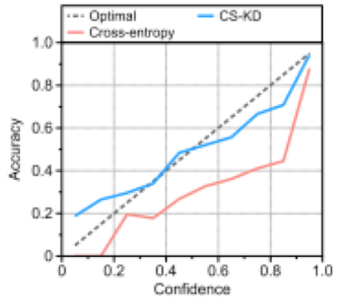
* **Bird** (CUB-200-2011; 200 labels), **dog** (Stanford dogs;120 labels), **indoor** (MIT67; 67 labels)

Experiments & Results

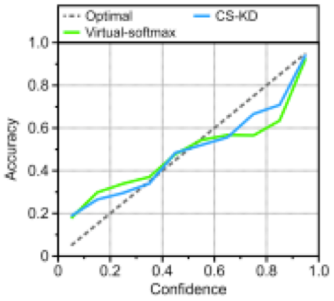
Calibration Effects

Measurement	Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
Top-5 ↓	Cross-entropy	6.91±0.09	22.21±0.29	22.30±0.68	11.80±0.27	19.25±0.53
	AdaCos	9.99±0.20	22.24±0.11	15.24±0.66	11.02±0.22	19.05±2.33
	Virtual-softmax	8.54±0.11	24.15±0.17	13.16±0.20	8.64±0.21	19.10±0.20
	Maximum-entropy	7.29±0.12	21.53±0.50	19.80±1.21	10.90±0.31	20.47±0.90
	Label-smoothing	7.18±0.08	20.74±0.31	22.40±0.85	13.41±0.40	19.53±0.75
	CS-KD (ours)	**5.69**±0.03	19.21±0.04	**13.07**±0.26	**8.55**±0.07	**17.46**±0.38
	CS-KD-E (ours)	5.93±0.06	**19.12**±0.34	13.74±0.91	8.57±0.13	18.21±0.45
ECE ↓	Cross-entropy	15.45±0.33	14.08±0.76	18.39±0.76	15.05±0.35	17.99±0.72
	AdaCos	73.76±0.35	55.09±0.41	63.39±0.06	65.38±0.33	54.00±0.52
	Virtual-softmax	8.02±0.55	4.60±0.67	11.68±0.66	7.91±0.38	11.21±1.00
	Maximum-entropy	56.41±0.36	42.68±0.31	50.52±1.20	51.53±0.28	42.41±1.74
	Label-smoothing	13.20±0.60	**2.67**±0.48	15.70±0.81	11.60±0.40	8.79±2.47
	CS-KD (ours)	5.17±0.40	7.26±0.93	15.44±0.92	10.46±1.08	15.56±0.29
	CS-KD-E (ours)	**4.69**±0.56	3.79±0.35	**8.75**±0.49	**4.70**±0.18	**8.06**±1.90
R@1 ↑	Cross-entropy	61.38±0.64	30.59±0.42	33.92±1.70	47.51±1.02	31.42±1.00
	AdaCos	67.95±0.42	44.66±0.52	54.86±0.24	58.37±0.43	42.39±1.91
	Virtual-softmax	68.35±0.48	44.69±0.58	55.56±0.74	59.71±0.56	44.20±0.90
	Maximum-entropy	**71.51**±0.29	39.18±0.79	48.66±2.10	60.05±0.45	38.06±3.32
	Label-smoothing	71.44±0.03	34.79±0.67	41.59±0.94	54.48±0.68	35.15±1.54
	CS-KD (ours)	71.15±0.15	**47.15**±0.40	**59.06**±0.38	**62.67**±0.07	**46.74**±1.48
	CS-KD-E (ours)	70.57±0.57	45.52±0.35	58.44±1.09	62.03±0.30	44.82±1.22

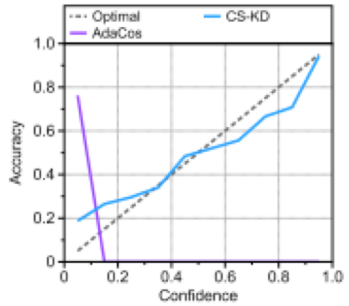
Table 6. Top-5 error, ECE, and Recall at 1 (R@1) rates (%) of ResNet-18 on various image classification tasks. We denote our method combined with the sample-wise regularization by CS-KD-E. The arrow on the right side of the evaluation metric indicates ascending or descending order of the value. We reported the mean and standard deviation over three runs with different random seeds, and the best results are indicated in bold.



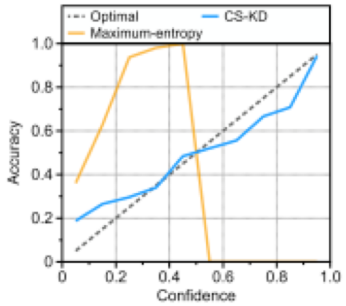
(a) Cross-entropy



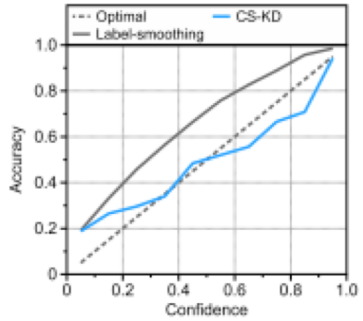
(b) Virtual-softmax



(c) AdaCos



(d) Maximum-entropy



(e) Label-smoothing

Conclusion

1. Discover **a simple regularization method** to enhance the generalization performance of DNN

2. It works by **penalizing the predictive distribution between different samples of the same label by minimizing the KL divergence**

3. CS-KD **regularizes the knowledge on wrong predictions** and encourages the model to produce more meaningful predictions

4. Demonstrate that CS-KD can be useful for the **generalization and calibration of neural networks**

Thank you