Regularizing Class-wise Predictions via Self-knowledge Distillation

Sukmin Yun[1*] Jongjin Park[1*] Kimin Lee[2†] Jinwoo Shin[1]
[1]Korea Advanced Institute of Science and Technology, South Korea
[2]University of California, Berkeley, USA
{sukmin.yun, jongjin.park, jinwoos}@kaist.ac.kr kiminlee@berkeley.edu

김 성 철

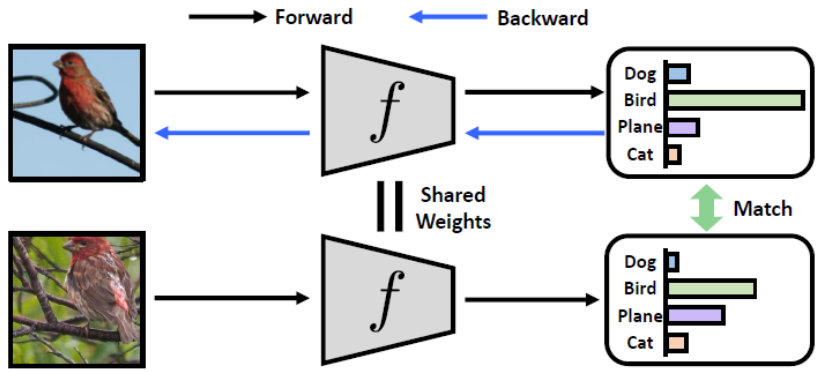
Contents

Introduction

- Regularization
  - 네트워크의 parameter가 많아질 수록 overfitting과 poor generalizatio이 발생
    - early stopping, L_1/L_2-regularization, dropout, batch normalization, data augmentation
  - **Regularizing the predictive distribution of DNNs**
    - Label-smoothing, entropy maximization, angular-margin
    - Network calibration, novelty detection, exploration in reinforcement learning에도 영향미침
- **Dark knowledge**
  - 잘못된 예측의 knowledge
  - Knowledge distillation에서 처음으로 중요성이 입증됨

Introduction

- **Class-wise Self-Knowledge Distillation (CS-KD)**
  - 같은 클래스의 다른 sample에 대한 predictive distribution을 matching or distilling
  - 같은 클래스의 sample들이 잘못된 예측을 하더라도 비슷한 예측을 하도록 유도
    - Predictive distribution의 일관성
  - **Preventing overconfident prediction & Reducing intra-class variation**
  - 다른 regularization 방법들보다 낮은 top-1 error rate
  - 더 좋은 top-5 error rate와 expected calibration error
  - 최근 self-distillation 방식들보다 좋은 top-1 error rate
  - Mixup, knowledge distillation 등 방법들과 합쳤을 때 더 좋은 성능



(a) Overview of our regularization scheme

Class-wise Self-Knowledge Diatillation

- Softmax classifier

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

- Class-wise regularization
  - 같은 클래스의 sample들에 대해 일정한 predictive distribution을 유도
  - **Class-wise regularization loss**

$$\mathcal{L}_{\text{cls}}(x,x';\theta,T) \coloneqq \text{KL}\left(P\left(y|x';\tilde{\theta},T\right)||P(y|x;\theta,T)\right)$$

    - x, x' : input, another randomly sampled input having the same label y
    - KL : the Kullback-Leibler divergence
    - $\tilde{\theta}$: a fixed copy of the parameters θ. **Stop gradient to avoid the model collapse issue**

Class-wise Self-Knowledge Distillation

- Class-wise regularization (Cont.)
  - Total training loss $\mathcal{L}_{\text{CS-KD}}$

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

    - $\mathcal{L}_{\text{CE}}$: the standard cross-entropy loss
    - $\lambda_{\text{cls}} > 0$: a loss weight for the class-wise regularization
    - Original KD와 동일하게 the square of the temperature T^2 적용

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T}\right)$$

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

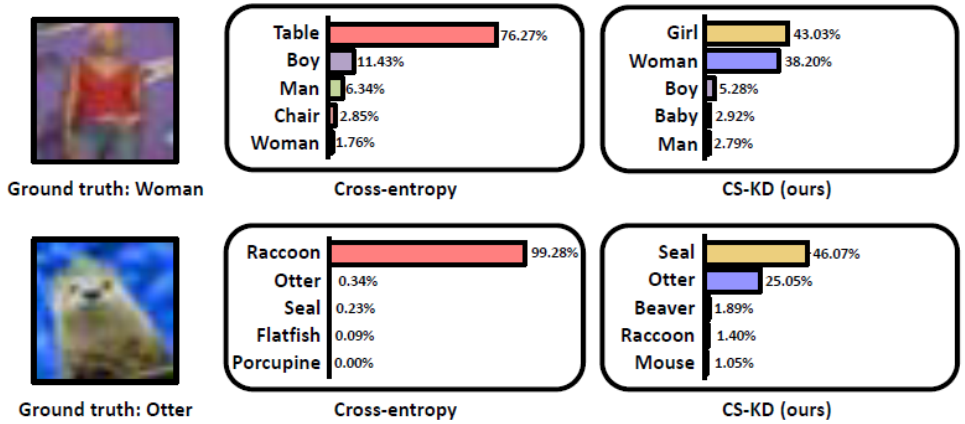
Algorithm 1 Class-wise self-knowledge distillation

Initialize parameters θ.
while θ has not converged **do**
  Sample a batch $(\mathbf{x}, y)$ from the training dataset.
  Sample another batch $\mathbf{x}'$ randomly, which has the same label y from the training dataset.
  Update parameters θ by computing the gradients of the proposed loss function $\mathcal{L}_{\text{CS-KD}}(\mathbf{x}, \mathbf{x}', y; \theta, T)$ in (1).
end while

https://arxiv.org/pdf/1503.02531.pdf

Class-wise Self-Knowledge Distillation

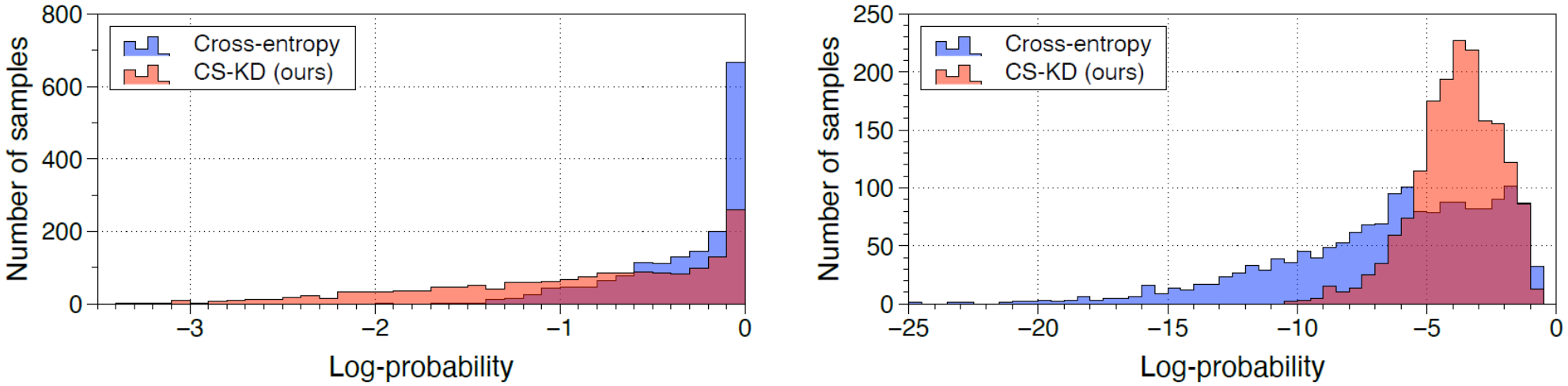
- Effects of class-wise regularization
  - **Preventing overconfident predictions**
    - 다른 sample들의 model-prediction을 soft-label로 사용
      - Label-smoothing보다 현실적 $y_k^{LS} = y_k(1-\alpha) + \alpha/K$
  - **Reducing the intra-class variations**
    - 같은 클래스의 두 logit 사이의 거리를 최소화
  - Softmax의 prediction value 조사
    - PreAct ResNet-18 trained on the CIFAR-100
    - CIFAR-100에서 잘못 예측한 데이터로 확인
    - Overconfident prediction 완화
    - Ground-truth 클래스의 prediction value 강화



(b) Top-5 softmax scores on misclassified samples

Class-wise Self-Knowledge Distillation

- Effects of class-wise regularization (Cont.)
  - Log-probabilities of the softmax scores
    - (a) 잘못 예측한 sample의 confident prediction이 낮음
    - (b) 잘못 예측한 sample의 ground-truth class의 score가 높음



(a) Log-probabilities of predicted labels on misclassified samples

(b) Log-probabilities of ground-truth labels on misclassified samples

Figure 2. Histogram of log-probabilities of (a) the predicted label, *i.e.*, top-1 softmax score, and (b) the ground-truth label on misclassified samples by networks trained by the cross-entropy (baseline) and CS-KD. The networks are trained on PreAct ResNet-18 for CIFAR-100.

Experiments

- Experimental setup
  - Datasets
    - CIFAR-100, TinyImageNet : datasets for conventional classification tasks
    - CUB-200-2011, Standford Dogs, MIT67 : datasets for fine-grained classification tasks
      - 시각적으로 유사한 클래스들이 존재, 클래스당 training sample이 적음
    - ImageNet : a large-scale classification task
  - Network architecture
    - ResNet-18 with 64 filters, DenseNet-121 with a growth rate of 32 : fine-grained classification
    - PreAct ResNet-18 : conventional classification
  - Hyper-parameters
    - SGD with momentum 0.9, weight decay 0.0001, an initial learning rate of 0.1 (divided by 10 after epochs 100 and 150)
    - 200 epochs / batch size : 128(conventional), 32(fine-grained) / Flipping, random cropping
    - T {1, 4} / λ_{cls} {1, 2, 3, 4}

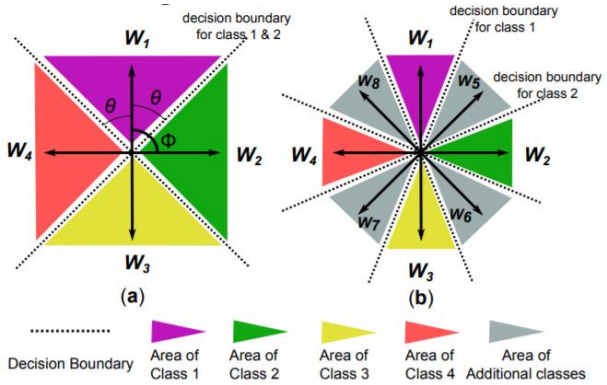
https://arxiv.org/pdf/1905.00292.pdf
https://arxiv.org/pdf/1811.12611.pdf
https://arxiv.org/pdf/1809.05934.pdf
https://arxiv.org/pdf/1906.02629.pdf
https://www.aaai.org/ojs/index.php/AAAI/article/view/4498
https://arxiv.org/pdf/1905.08094.pdf

Experiments

- Experimental setup (Cont.)
  - Baselines

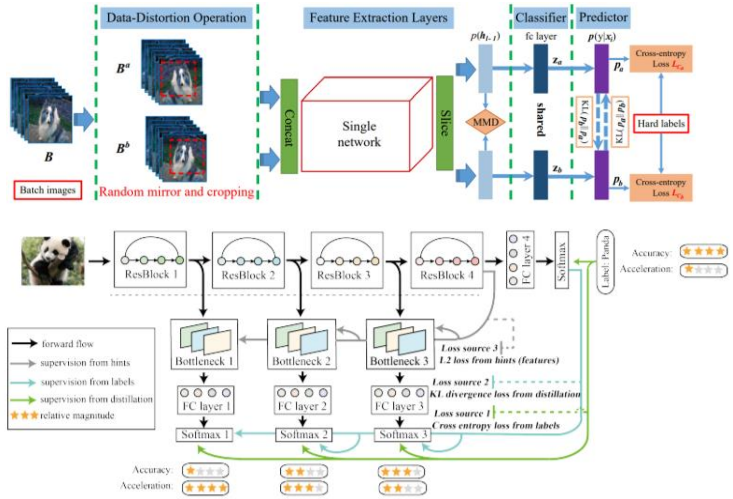
- **AdaCos** [58].[5] AdaCos dynamically scales the cosine similarities between training samples and corresponding class center vectors to maximize angular-margin.
- **Virtual-softmax** [5]. Virtual-softmax injects an additional virtual class to maximize angular-margin.
- **Maximum-entropy** [13, 36]. Maximum-entropy is a typical entropy regularization, which maximizes the entropy of the predictive distribution.
- **Label-smoothing** [32, 43]. Label-smoothing uses soft labels that are a weighted average of the one-hot labels and the uniform distribution.
- **DDGSD** [53]. Data-distortion guided self-distillation (DDGSD) is one of the consistency regularization techniques, which forces the consistent outputs across different augmented versions of the data.
- **BYOT** [57]. Be Your Own Teacher (BYOT) transfers the knowledge in the deeper portion of the networks into the shallow ones.

$$P_{i,j} = \frac{e^{\tilde{s}\cdot\cos\theta_{i,j}}}{\sum_{k=1}^{C} e^{\tilde{s}\cdot\cos\theta_{i,k}}}$$



$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \widehat{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} \left[\mathbb{D}_{\mathsf{KL}}\left(\bar{\mathbf{y}}(\mathbf{x})||p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\right) - \gamma \mathsf{H}[p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})]\right]$$

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$



Experiments

- Experimental setup (Cont.)
  - Evaluation metric

- **Top-1 / 5 error rate.** The top-k error rate is the fraction of test samples for which the correct label is not in the top-k confidences. We measure top-1 and top-5 error rates to evaluate the generalization performances.
- **Expected Calibration Error (ECE).** ECE [16, 33] approximates the difference in expectation between confidence and accuracy. It is calculated by partitioning predictions into M equally-spaced bins and taking a weighted average of bins' difference of confidence and accuracy, *i.e.*, ECE $= \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$, where n is the number of samples, B_m is the set of samples whose confidence falls into the m-th interval, and $\mathrm{acc}(B_m)$, $\mathrm{conf}(B_m)$ are the accuracy and the average confidence of B_m, respectively. We measure ECE with 20 bins to evaluate whether the model represents the true correctness likelihood.
- **Recall at k (R@k).** Recall at k is the percentage of test samples that have at least one from the same class in k nearest neighbors on the feature space. To measure the distance between two samples, we use L_2-distance between their pooled features of the penultimate layer. We compare the recall at $k = 1$ scores to evaluate intra-class variations of learned features.

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i),$$

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i,$$

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \Big|,$$

The average confidence of LeNet closely matches its accuracy, while the average confidence of the ResNet is substantially higher than its accuracy. This is further illustrated in the bottom row reliability diagrams (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005), which show accuracy as a function of confidence. We see that LeNet is well-calibrated, as confidence closely approximates the expected accuracy (i.e. the bars align roughly along the diagonal). On the other hand, the ResNet's accuracy is better, but does not match its confidence.

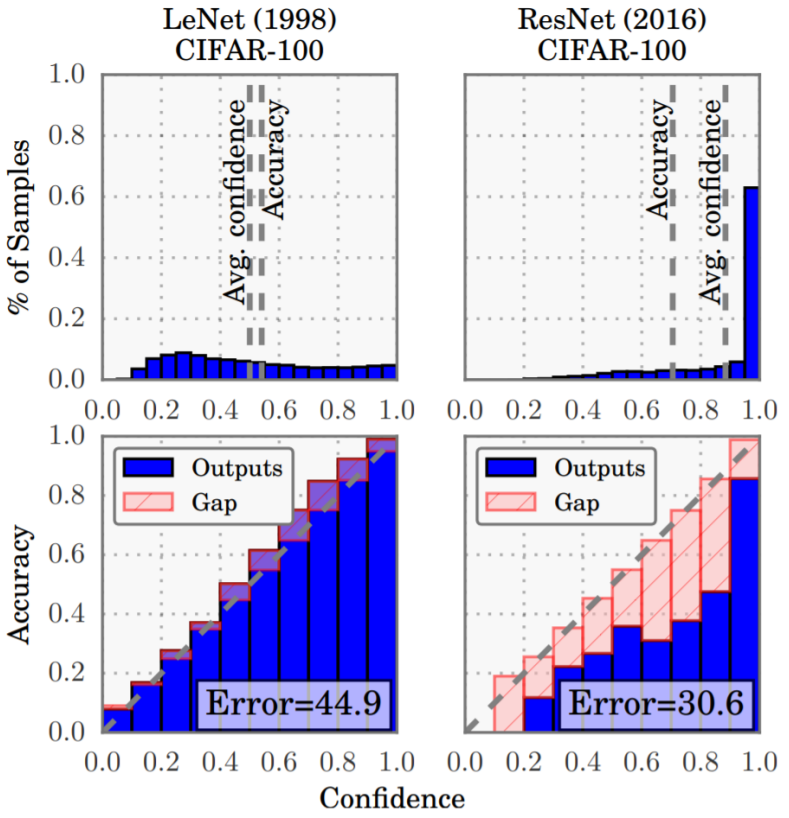


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

https://arxiv.org/pdf/1706.04599.pdf

Experiments

- Classification accuracy
  - 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.

Measurement	Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
Top-5 ↓	Cross-entropy	6.91$_{\pm 0.09}$	22.21$_{\pm 0.29}$	22.30$_{\pm 0.68}$	11.80$_{\pm 0.27}$	19.25$_{\pm 0.53}$
	AdaCos	9.99$_{\pm 0.20}$	22.24$_{\pm 0.11}$	15.24$_{\pm 0.66}$	11.02$_{\pm 0.22}$	19.05$_{\pm 2.33}$
	Virtual-softmax	8.54$_{\pm 0.11}$	24.15$_{\pm 0.17}$	13.16$_{\pm 0.20}$	8.64$_{\pm 0.21}$	19.10$_{\pm 0.20}$
	Maximum-entropy	7.29$_{\pm 0.12}$	21.53$_{\pm 0.50}$	19.80$_{\pm 1.21}$	10.90$_{\pm 0.31}$	20.47$_{\pm 0.90}$
	Label-smoothing	7.18$_{\pm 0.08}$	20.74$_{\pm 0.31}$	22.40$_{\pm 0.85}$	13.41$_{\pm 0.40}$	19.53$_{\pm 0.75}$
	CS-KD (ours)	**5.69**$_{\pm 0.03}$	19.21$_{\pm 0.04}$	**13.07**$_{\pm 0.26}$	**8.55**$_{\pm 0.07}$	**17.46**$_{\pm 0.38}$
	CS-KD-E (ours)	5.93$_{\pm 0.06}$	**19.12**$_{\pm 0.34}$	13.74$_{\pm 0.91}$	8.57$_{\pm 0.13}$	18.21$_{\pm 0.45}$

Experiments

- Classification accuracy
  - 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	$23.85_{\pm 1.57}$	$\mathbf{41.48_{\pm 0.12}}$	$41.17_{\pm 1.28}$	$31.53_{\pm 0.54}$	$41.17_{\pm 2.46}$
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)	$\mathbf{21.99_{\pm 0.13}}$ **(-11.0%)**	$41.62_{\pm 0.38}$ (- 4.4%)	$\mathbf{33.28_{\pm 0.99}}$ **(-27.7%)**	$\mathbf{30.85_{\pm 0.28}}$ **(-15.0%)**	$\mathbf{40.45_{\pm 0.45}}$ **(- 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

- Classification accuracy
  - 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

- Classification accuracy
  - 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

- Ablation study
  - Feature embedding analysis

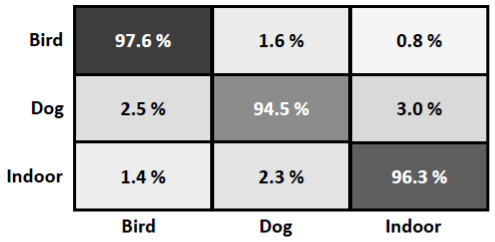
(a) Cross-entropy (b) Virtual-softmax (c) AdaCos (d) CS-KD (ours)

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.

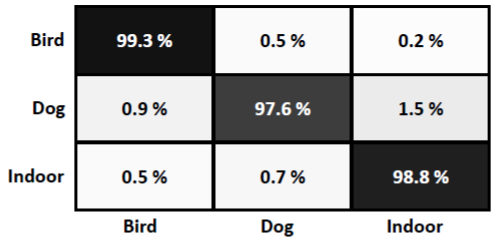
Measurement	Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
R@1 ↑	Cross-entropy	61.38±0.64	30.59±0.42	33.92±1.70	47.51±1.02	31.42±1.00
	AdaCos	67.95±0.42	44.66±0.52	54.86±0.24	58.37±0.43	42.39±1.91
	Virtual-softmax	68.35±0.48	44.69±0.58	55.56±0.74	59.71±0.56	44.20±0.90
	Maximum-entropy	**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

Experiments

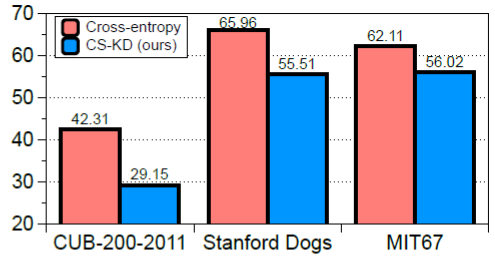
- Ablation study
  - Hierarchical image classification
    - 387 fine-grained labels and three hierarchy labels : bird(CUB-200-2011), dog(Standford Dogs), indoor (MIT67)
    - Fine-grained label당 30 sample씩 임의 추출 후 학습. Original testset으로 테스트
    - Fine-grained label을 예측하고 hierarchical classification accuracy 측정



(a) Cross-entropy



(b) CS-KD (ours)



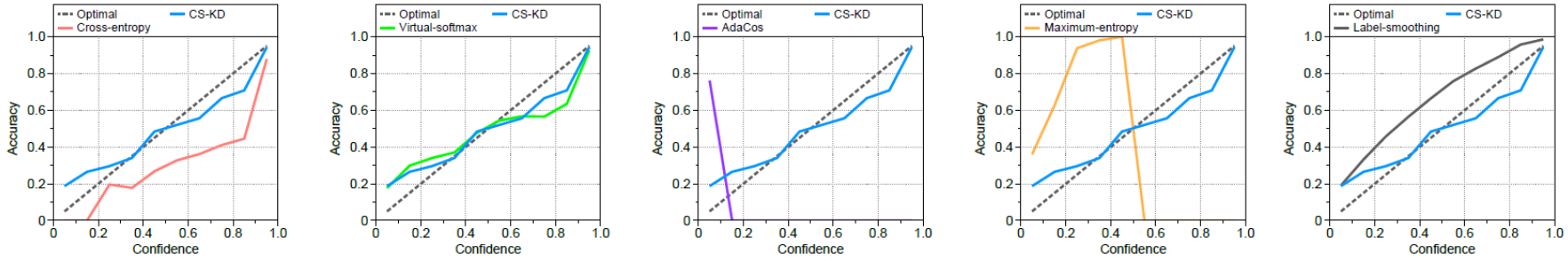
(c) Top-1 error rates (%)

CUB-200-2011	Stanford Dogs	MIT67
$46.00_{\pm 1.43}$	$36.29_{\pm 0.32}$	$44.75_{\pm 0.80}$
$35.47_{\pm 0.07}$	$32.66_{\pm 0.34}$	$42.66_{\pm 0.43}$
$35.03_{\pm 0.51}$	$31.48_{\pm 0.16}$	$42.86_{\pm 0.71}$
$39.86_{\pm 1.11}$	$32.41_{\pm 0.20}$	$43.36_{\pm 1.62}$
$42.99_{\pm 0.99}$	$35.30_{\pm 0.66}$	$44.40_{\pm 0.71}$
$33.28_{\pm 0.99}$ (-27.7%)	**$30.85_{\pm 0.28}$ (-15.0%)**	**$40.45_{\pm 0.45}$ (- 9.6%)**

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.

Experiments

- Calibration effects
  - Plotted identity function (dashed diagonal) : perfect calibration



(a) Cross-entropy (b) Virtual-softmax (c) AdaCos (d) Maximum-entropy (e) Label-smoothing

Figure 5. Reliability diagrams [9, 34] show accuracy as a function of confidence, for PreAct ResNet-18 trained on CIFAR-100 using (a) Cross-entropy, (b) Virtual-softmax, (c) AdaCos, (d) Maximum-entropy, and (e) Label-smoothing. All methods are compared with our proposed method, CS-KD. Perfect calibration [16] is plotted by dashed diagonals (Optimal) for all.

Measurement	Method	CIFAR-100	TinyImageNet	CUB-200-2011	Stanford Dogs	MIT67
ECE ↓	Cross-entropy	15.45$_{\pm 0.33}$	14.08$_{\pm 0.76}$	18.39$_{\pm 0.76}$	15.05$_{\pm 0.35}$	17.99$_{\pm 0.72}$
	AdaCos	73.76$_{\pm 0.35}$	55.09$_{\pm 0.41}$	63.39$_{\pm 0.06}$	65.38$_{\pm 0.33}$	54.00$_{\pm 0.52}$
	Virtual-softmax	8.02$_{\pm 0.55}$	4.60$_{\pm 0.67}$	11.68$_{\pm 0.66}$	7.91$_{\pm 0.38}$	11.21$_{\pm 1.00}$
	Maximum-entropy	56.41$_{\pm 0.36}$	42.68$_{\pm 0.31}$	50.52$_{\pm 1.20}$	51.53$_{\pm 0.28}$	42.41$_{\pm 1.74}$
	Label-smoothing	13.20$_{\pm 0.60}$	**2.67**$_{\pm \mathbf{0.48}}$	15.70$_{\pm 0.81}$	11.60$_{\pm 0.40}$	8.79$_{\pm 2.47}$
	CS-KD (ours)	5.17$_{\pm 0.40}$	7.26$_{\pm 0.93}$	15.44$_{\pm 0.92}$	10.46$_{\pm 1.08}$	15.56$_{\pm 0.29}$
	CS-KD-E (ours)	**4.69**$_{\pm \mathbf{0.56}}$	3.79$_{\pm 0.35}$	**8.75**$_{\pm \mathbf{0.49}}$	**4.70**$_{\pm \mathbf{0.18}}$	**8.06**$_{\pm \mathbf{1.90}}$

Experiments

- Calibration effects (Cont.)
  - Consistency loss와 결합 $\mathcal{L}_{\text{CS-KD-E}}$

$$\mathcal{L}_{\text{CS-KD-E}}(x, x', y; \theta, T) \coloneqq \mathcal{L}_{\text{CS-KD}}\left(x_{\text{aug}}, x'_{\text{aug}}, y; \theta, T\right) + \lambda_{\text{E}} \cdot T^2 \cdot \text{KL}\left(P\left(y|x; \tilde{\theta}, T\right)||P\left(y|x_{\text{aug}}; \theta, T\right)\right)$$

    - x_{aug} : an augmented sample that is generated by the data augmentation technique
    - $\lambda_{\text{E}} > 0$: the loss weight for balancing

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

Conclusion

- A simple regularization method to enhance the generalization performance of DNN
  - 같은 label의 다른 sample들의 predictive distribution의 Kullback-Leibler divergence를 최소화
- Generalization and calibration of neural network
- Applicable with a broader range of applications
  - Exploration in deep reinforcement learning
  - Transfer learning
  - Face verification
  - Detection of out-of-distribution samples

감 사 합 니 다