NEURAL INFORMATION PROCESSING SYSTEMS

Decoupled Kullback-Leibler Divergence Loss

Jiequan Cui[1], Zhuotao Tian[2], Zhisheng Zhong[3], Xiaojuan Qi[4], Bei Yu[3], Hanwang Zhang[1]

NTU[1] HIT(SZ)[2] CUHK[3] HKU[4]

Decoupled Kullback-Leibler Divergence Loss

Decoupled Kullback-Leibler Divergence Loss

- Kullback-Leibler (KL) Divergence Loss
  - Definition

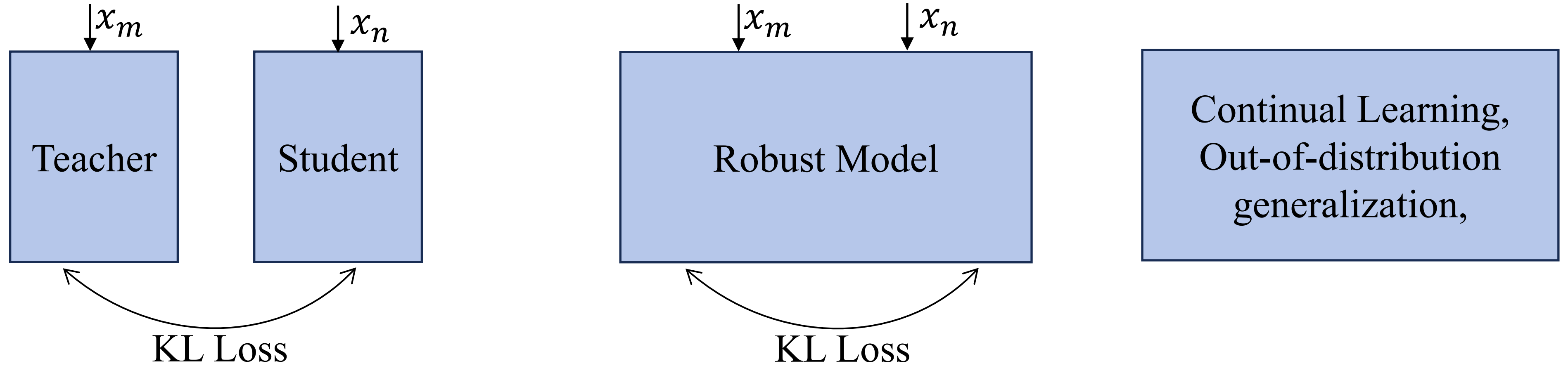
Assume $x_m, x_n \in X$, the KL loss encourages outputs consistency:

$$\mathcal{L}_{KL}(x_m, x_n) = \sum_{j=1}^{C} s_m^j \log \frac{s_m^j}{s_n^j}$$

where o_m, o_n are logit outputs, s_m, s_n are *softmax* scores.

Decoupled Kullback-Leibler Divergence Loss

- Kullback-Leibler (KL) Divergence Loss
  - Application



Knowledge Distillation

Adversarial Training

Others

Decoupled Kullback-Leibler Divergence Loss

- Kullback-Leibler (KL) Divergence Loss
  - Gradient Optimization

$$\frac{\partial \mathcal{L}_{KL}}{\partial o_m} = \sum_{k=1}^{C} ((\Delta m_{j,k} - \Delta n_{j,k}) * (s_m^k s_m^j))$$

$$\frac{\partial \mathcal{L}_{KL}}{\partial o_n} = s_m^j * \left(s_n^j - 1\right) + s_n^j * (1 - s_m^j)$$

where $\Delta m_{j,k} = o_m^j - o_m^k$ and $\Delta n_{j,k} = o_n^j - o_n^k$

Decoupled Kullback-Leibler Divergence Loss

- Decoupled Kullback-Leibler (DKL) Divergence Loss

Theorem 1. From the perspective of gradient optimization, the KL Divergence loss is equivalent to the following DKL Divergence loss when $\alpha = 1$ and $\beta = 1$.

$$\mathcal{L}_{DKL}(x_m, x_n) = \underbrace{\frac{\alpha}{4} \|\sqrt{S(w_m)}(\Delta_m - S(\Delta_n))\|^2}_{\textbf{weighted MSE (wMSE)}} - \underbrace{\beta \cdot S(s_m^T) \log s_n}_{\textbf{Cross-Entropy}}$$

where S(·) means stop gradients operation, $\mathrm{s}_\mathrm{m}^\mathrm{T}$ is the transpose of s_m, $\Delta m_{j,k} = o_m^j - o_m^k$ and $\Delta n_{j,k} = o_n^j - o_n^k$, summation is used for reduction of $\|\cdot\|^2$.

Decoupled Kullback-Leibler Divergence Loss

- Decoupled Kullback-Leibler (DKL) Divergence Loss
  - Significance of DKL Loss
    - Theorem 1. precisely reveals the relationships between KL, MSE, and Cross-Entropy losses
    - The gradients regrading o_m and o_n are asymmetric
    - The component of **wMSE** depends on sample-wise prediction scores which might suffer from biases.

$$\boxed{\mathcal{L}_{DKL}(x_m, x_n) = \underbrace{\frac{\alpha}{4} ||\sqrt{S(w_m)}(\Delta_m - S(\Delta_n))||^2}_{\text{weighted MSE (wMSE)}} - \underbrace{\beta \cdot S(s_m^T) \log s_n}_{\text{Cross-Entropy}}}$$

NEURAL INFORMATION PROCESSING SYSTEMS

Decoupled Kullback-Leibler Divergence Loss

- Decoupled Kullback-Leibler (DKL) Divergence Loss
  - Improvements of DKL Loss

Symmetric gradients on o_m and o_n

$$\mathcal{L}_{DKL-KD}(x_m, x_n) = \underbrace{\frac{\alpha}{4}||\sqrt{S(w_m)}(\Delta_m - \Delta_{\boldsymbol{n}})||^2}_{\textbf{weighted MSE (wMSE)}} - \underbrace{\beta \cdot S(s_m^T)\log s_n}_{\textbf{Cross-Entropy}}$$

Introduce global information on **wMSE**

$$\bar{w}_y^{j,k} = \bar{s}_y^j * \bar{s}_y^k, \text{where y is ground-truth label of } x_m,\ \bar{s}_y = \frac{1}{|X_y|}\sum_{x_i \in X_y} s_i$$

$$\mathcal{L}_{IKL}(x_m, x_n) = \underbrace{\frac{\alpha}{4}||\sqrt{S(\bar{\boldsymbol{w}}_{\boldsymbol{m}})}(\Delta_m - \Delta_{\boldsymbol{n}})||^2}_{\textbf{weighted MSE (wMSE)}} - \underbrace{\beta \cdot S(s_m^T)\log s_n}_{\textbf{Cross-Entropy}}$$

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

1920 HIT

Decoupled Kullback-Leibler Divergence Loss

Experimental Results

Adversarial Training

New state-of-the-art on Auto-Attack benchmark

Table 2: **Test accuracy (%) of clean images and robustness (%) under AutoAttack on CIFAR-100.** All results are the average over three trials.

Dataset	Method	Architecture	Augmentation Type	Clean	AA
CIFAR-100 (ℓ_∞, $\epsilon = 8/255$)	AWP	WRN-34-10	Basic	60.38	28.86
	LBGAT	WRN-34-10	Basic	60.64	29.33
	LAS-AT	WRN-34-10	Basic	64.89	30.77
	ACAT	WRN-34-10	Basic	65.75	30.23
	IKL-AT	WRN-34-10	Basic	**65.76**	**31.91**
	ACAT	WRN-34-10	AutoAug	**68.74**	31.30
	IKL-AT	WRN-34-10	AutoAug	66.08	**32.53**
	DM-AT [39]	WRN-28-10	50M Generated Data	72.58	38.83
	IKL-AT	WRN-28-10	50M Generated Data	**73.85**	**39.18**

NEURAL INFORMATION PROCESSING SYSTEMS

Decoupled Kullback-Leibler Divergence Loss

Experimental Results

Knowledge Distillation

Competitive performance on knowledge distillation

Table 4: **Top-1 accuracy (%) on the ImageNet validation and training speed (sec/iteration) comparisons.** Training speed is calculated on 4 Nvidia GeForce 3090 GPUs with a batch of 512 224x224 images. All results are the average over three trials.

Distillation Manner	Teacher / Student	Extra Parameters	ResNet34 73.31 / ResNet18 69.75		ResNet50 76.16 / MobileNet 68.87	
Features	AT	✗	70.69		69.56	
	OFD	✔	70.81		71.25	
	CRD	✔	71.17		71.37	
	ReviewKD	✔	71.61	0.319 s/iter	72.56	0.526 s/iter
Logits	DKD	✗	71.70		72.05	
	KD	✗	71.03		70.50	
	IKL-KD	✗	**71.91**	**0.197 s/iter**	**72.84**	**0.252 s/iter**

Table 5: **Peformance (%) on imbalanced data, *i.e.*, the ImageNet-LT.**

Method	Teacher	Student	Many(%)	Medium(%)	Few(%)	All(%)
Baseline	-	ResNet-18	63.16	33.47	5.88	41.15
Baseline	-	ResNet-50	67.25	38.56	8.21	45.47
Baseline	-	ResNet-101	68.91	42.32	11.24	48.33
KL-KD	ResNeXt-101	ResNet-18	64.6	37.88	9.53	44.32
KL-KD	ResNeXt-101	ResNet-50	68.83	42.31	11.37	48.31
IKL-KD	ResNeXt-101	ResNet-18	66.60	38.53	8.19	**45.21**
IKL-KD	ResNeXt-101	ResNet-50	70.06	43.47	10.99	**49.29**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NEURAL INFORMATION PROCESSING SYSTEMS

Decoupled Kullback-Leibler Divergence Loss

Code

Paper

Thank You!

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE